2015

Predicting ESL learners' oral proficiency by measuring the collocations in their spontaneous speech

Jing Xu
Iowa State University

Predicting ESL learners' oral proficiency by measuring the collocations in their spontaneous speech

by

Jing Xu

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Applied Linguistics and Technology

Program of Study Committee:
Carol A. Chapelle, Major Professor
Dan Douglas
Volker Hegelheimer
Jean Goodwin
Zhengyuan Zhu

Iowa State University

Ames, Iowa

2015

Dedication

To my beloved wife Hong Wang and parents Guanghong Xu and Ruiying Li

TABLE OF CONTENTS

LIST OF TABLES vi
LIST OF FIGURES ix
ACKNOWLEDGEMENT x
ABSTRACT xiii
CHAPTER 1: INTRODUCTION 1
1.1 Automated Speech Evaluation (ASE): Is It a Fairy Tale? 1
1.2 Background of the Study 2
1.2.1 Restricted Automated Speech Evaluation (RASE) 4
1.2.2 Unrestricted Automated Speech Evaluation (UASE) 8
1.2.3 The Great Promise of UASE 10
1.3 Statement of the Problem 11
1.3.1 What is an Interpretive/Use Argument? 12
1.3.2 A Hypothetical Interpretive/Use Argument for a Semi-Automated ITA Oral English Assessment 12
1.3.3 Collocation and Automated Speech Evaluation 19
1.3.4 Inspiration from Previous Research 21
1.4 Goals and Significance of the Study 22
1.5 Outline of Dissertation 23
CHAPTER 2: LITERATURE REVIEW 25
2.1 Defining the Construct of Collocation 25
2.2 Collocational Knowledge and General L2 Proficiency 29
2.3 Collocation and Speaking 33
2.3.1 Bygate's Model of Speaking as a Process 33
2.3.2 Levelt's Modular Model of Speech Production 38
2.3.3 De Bot's Bilingual Production Model 45
2.3.4 Kormos' Model of Bilingual Speech Production 48
2.3.5 Logical Analysis of the Relationship between Collocation and L2 Oral Proficiency 52
2.4 The Construct of Spoken Collocational Competence 55
2.5 Automated Evaluation of Spontaneous L2 Speech 57
2.5.1 The Exciting Prospect of ASE 57
2.5.2 The Architecture of SpeechRater 59
2.5.3 The Role of Construct Theory in Developing SpeechRater 64

2.6 Technology of Automated Collocation Extraction and Evaluation 65
2.7 Chapter Summary 67
2.8 Research Questions 68
CHAPTER 3: METHODOLOGY 70
3.1 Overall Research Design 70
3.2 Data Collection 71
3.2.1 Data Source 72
3.2.2 Sample Selection 75
3.2.3 Digitization and Transcription 76
3.3 Data Annotation 78
3.3.1 Target Collocations 78
3.3.2 Collocation Extraction and Validation 80
3.3.3 Collocation Coding Schemes 83
3.3.4 Procedure of Collocation Coding 90
3.4 Data Management and Analysis 92
3.4.1 Evaluation of Manual Collocation Extraction 93
3.4.2 Development of Collocation Measures 93
3.4.3 Reliabilities of Collocation Coding and Oral Proficiency Measures 98
3.4.4 RQ1: The Characteristics of the Collocation Occurrences in Learner Speech 98
3.4.5 RQ2: Variation of the Collocation Measures among Proficiency Level Groups 99
3.4.6 RQ3: Prediction of Oral Proficiency based on Collocation Measurement 100
3.4.7 RQ4: The Context Effect on Prediction 102
3.5 Chapter Summary 103
CHAPTER 4: RESULTS 104
4.1 The Characteristics of the Collocation Occurrences in Learner Speech 104
4.1.1 Reliability Estimates of Collocation Extraction and Coding 104
4.1.2 Descriptive Statistics 106
4.1.3 Wilcox Rank Test 109
4.2 Variation of the Collocation Measures among Proficiency Level Groups 110
4.2.1 Measures on Semantic Accuracy 111
4.2.2 Measures on Grammatical Accuracy 113
4.2.3 Measures on Restrictedness or Precision 116
4.2.4 Measures on Transparency 117
4.2.5 Measures on Automaticity 118

4.2.6 The Composite Measure 119
4.2.7 A Summary of Promising Collocation Measures 121
4.3 Prediction of Oral Proficiency based on Collocation Measurement 121
4.3.1 Correlational Analyses 121
4.3.2 Multiple Regression for Predicting SPEAK Score 126
4.3.3 The Most Parsimonious Regression Model for Predicting SPEAK Score 130
4.3.4 Multiple Regression for Predicting TEACH Score 132
4.3.5 The Most Parsimonious Regression Model for Predicting TEACH Score 135
4.3.6 Logistic Regression 136
4.3.7 Composite Scoring vs. Regression Models 138
4.4 Cross-validation across Two Speaking Contexts 138
4.5 Chapter Summary 139
CHAPTER 5: CONCLUSIONS AND DISCUSSION 141
5.1 A Summary and Discussion of the Major Findings 141
5.1.1 Basic Characteristics of Collocations 141
5.1.2 Differences among Proficiency Level Groups 145
5.1.3 Prediction for L2 Oral Proficiency 147
5.1.4 Context Effect on Prediction 149
5.2 Implications 150
5.2.1 Implications for L2 Speaking Theory 150
5.2.2 Implications for Automated Speech Evaluation 151
5.2.3 Implications for Training Chinese ITAs 152
5.3 Limitations and Directions for Future Research 154
5.4 Concluding Remarks 157
REFERENCES 158
APPENDIX A: INSTITUTIONAL REVIEW BOARD APPROVAL 176

LIST OF TABLES

Table 1.1 A Summary of the Warrants, Assumptions, and Potential Rebuttals Associated with the Evaluation, Explanation, Utilization, and Ramification Inferences (adapted from Chapelle, et al., 2008, 2010) 17

Table 2.1 A Summary of the Studies on the Relationship between L2 Collocational Knowledge and General L2 Proficiency 32

Table 2.2 Some Representative Candidate Features of SpeechRater 63

Table 3.1 An Overview of the Structure of the SPEAK and TEACH Exams 73

Table 3.2 Transcription Scheme Adapted from Du Bois (1991) 77

Table 3.3 Ten Target Collocation Patterns 80

Table 3.4 The Coding Scheme of Semantic Accuracy 85

Table 3.5 A Summary of Surface Error Types 86

Table 3.6 The Coding Scheme for Transparency 88

Table 3.7 The Coding Scheme for Restrictedness 88

Table 3.8 The Coding Scheme for Automaticity 89

Table 3.9 A Summary of the Coded Collocation Features 89

Table 3.10 A Summary of Operational Collocation Performance Measures (OCPMs) 97

Table 4.1 Inter-coder Agreement by Collocation Coding Categories (n = 60) 106

Table 4.2 Descriptive Statistics of the Coded Collocations by Coding Categories (n = 60) 107

Table 4.3 Descriptive Statistics of Frequency and Semantic Accuracy of Extracted Collocations by Syntactic Pattern (n = 60) 107

Table 4.4 Descriptive Statistics of Collocation Performance Measures (n=60) 108

Table 4.5 Tests of Normality of the ACP_OK, ACP_ERR, and ACP_RAT in the SPEAK and TEACH Datasets (n=60) 110

Table 4.6 Tests of Normality of the ACP_OK, ACP_ERR, and ACP_RAT Scores in Each Oral Proficiency Level (n=60) 112

Table 4.7 Descriptive Statistics of ACP_OK, ACP_ERR, and Inv(ACP_RAT) Scores across Oral Proficiency Levels (n=60) 113

Table 4.8 Tests of Normality of the GRA_OK, GRA_ERR, and GRA_RAT Scores in Each Proficiency Level (n=60) 114

Table 4.9 Descriptive Statistics of GRA_OK, GRA_ERR, and GRA_RAT Scores across Proficiency Level Groups (n=60) 115

Table 4.10 Tests of Normality of RES_FRE and RES_PRO Scores in Each Oral Proficiency Level (n=60) 116

Table 4.11 Descriptive Statistics of RES_FRE and RES_PRO Scores across Proficiency Level Groups (n=60) 117

Table 4.12 Tests of Normality of TRAN Scores in Each Proficiency Level (n=60) 118

Table 4.13 Descriptive Statistics of TRAN Scores across Proficiency Level Groups (n=60) 118

Table 4.14 Tests of Normality of CHOP Scores in Each Proficiency Level (n=60) 119

Table 4.15 Descriptive Statistics of CHOP Scores across Proficiency Level Groups (n=60) 119

Table 4.16 Tests of Normality of CPR Scores in Proficiency Level Groups (n=60) 120

Table 4.17 Descriptive Statistics of CPR Scores across Proficiency Level Groups (n=60). 120

Table 4.18 Pearson and Correlations between OCPMs and SPEAK/TEACH Scores, Placement and Speech Lengths (n=60) 125

Table 4.19 Collinearity Statistics for the SPEAK Dataset (n=60) 127

Table 4.20 Mean, Standard Deviation, and Intercorrelations for SPEAK Score and Predictor Variables (n=60) 129

Table 4.21 Regression Analysis Summary of a Full Model for Predicting SPEAK Score (n=60) 130

Table 4.22 Statistics for SPEAK Prediction Model Comparisons 131

Table 4.23 Regression Analysis Summary of the Most Parsimonious Model for Predicting SPEAK Score (n=60) 131

Table 4.24 Collinearity Statistics for the SPEAK Dataset (n=60)..... 132

Table 4.25 Mean, Standard Deviation, and Intercorrelations for TEACH Score and Predictor Variables (n=60)..... 134

Table 4.26 Regression Analysis Summary of a Full Model for Predicting TEACH Score (n=60)..... 135

Table 4.27 Statistics for TEACH Prediction Model Comparisons (n=60) 136

Table 4.28 Regression Analysis Summary of the Most Parsimonious Model for TEACH Score (n=60) 136

Table 4.29 Logistic Regression Analysis Summary for Predicting Certification Decisions (n=60)..... 137

Table 4.30 A Comparison of the Predictive Values of the Composite Scoring Approach and Regression Approach (n=60) 138

LIST OF FIGURES

Figure 1.1. A simple illustration of the construct coverage of UASE and RASE 3

Figure 1.2. The interpretive/use argument for using automated score for adjudication in the case of large discrepancies between two human ratings 14

Figure 2.1. The continuum of idiomaticity (Adapted from Howarth, 1998, p. 28) 28

Figure 2.2. A model of speaking as a process (Bygate, 1987, p. 50) 38

Figure 2.3. A modular model of speech production (Levelt, 1999b, p. 87) 41

Figure 2.4. A bilingual production model based on de Bot's (1992) description 47

Figure 2.5. A model of bilingual speech production (Kormos, 2006, p. 168) 50

Figure 2.6. The architecture of the SpeechRater (adapted from Xi, et al., 2008, p. 19) 60

Figure 2.7. The construct coverage of SpeechRater (adapted from Xi, et al., 2008, p. 29) 61

Figure 3.1. An illustration of the three phases of the dissertation study 71

Figure 3.2. The screenshot of a collocation coding spreadsheet 82

Figure 3.3. Three dimensions of collocation measurement 85

Figure 4.1. The scatterplots of OCPMs and SPEAK scores 123

Figure 4.2. The scatterplots of OCPMs and TEACH scores 124

Figure 4.3. A scatterplot of standardized residuals against standardized predicted SPEAK scores 128

Figure 4.4. A scatterplot of standardized residuals against standardized predicted TEACH scores 133

ACKNOWLEDGEMENT

My first and deepest thanks go to Professor Carol Chapelle, chair of my dissertation committee, who fostered and saw my growth into a language tester. She introduced me to the field of language testing, took me under her wing when I blew my first conference presentation, taught me how to write and publish academic papers, and walked me through the dissertation journey. She had tremendous influence on me but also gave me complete freedom in my designing of this dissertation study. When I 'hit a wall' or had self-doubt, she never lost trust in me and kept pushing me forward. Her critical feedback broadened my vision and shaped my thinking. I was so blessed to be able to work with her.

I am also deeply indebted to Professor Dan Douglas, a prominent language tester, with whom I took two courses: English for specific purposes and computer-assisted language testing. He is so personable that I learned from him many anecdotes about language testers in his generation. I came to realize that being a language tester is a fun job. He also provided me with insightful feedback throughout my dissertation journey. He and his lovely wife Ms. Felicity Douglas even helped me code a large proportion of my pilot study data which was such tedious work. My experience of driving with him from Ames, Iowa to Ann Arbor, Michigan for the Language Testing Research Colloquium (LTRC) in 2011 is memorable.

My appreciation goes out to Dr. Xiaoming Xi, Senior Research Director at Educational Testing Service. She was my mentor during my summer internship at the research institute in 2009. It was she who put my hands on a SpeechRater research

project and stirred up my strong interest in automated scoring of learner speech. This dissertation study was inspired by and an extension of my internship research.

I would like to express my sincere gratitude to Professor Volker Hegelheimer, Professor Jean Goodwin, and Professor Zhengyuan Zhu, who served as my committee members. They were constantly encouraging and supportive and their constructive feedback further improved the quality of this paper.

I also owe thanks to Professor Elena Cotos and the Academic Communication Program (formerly the SPEAK/TEACH program) for permission to use the SPEAK and TEACH exam data for conducting my dissertation research. Elena was my supervisor and colleague at Graduate College. She is incredibly hard-working and has set a role model for me. She was always supportive to my professional development.

Thanks to Professor Gary Ockey for providing me with valuable suggestions at my planning stage of dissertation. We also sat together in Professor Chapelle's validity theory seminar. Our exchange of ideas on the evolution of validity theory in the seminar inspired a paper of mine to be published in the *Language Testing* journal.

I would like to thank Educational Testing Service for sponsoring my dissertation research through the TOEFL Small Grants for Doctoral Research in Second or Foreign Language Assessment. Their generous financial support allowed me to get data coding done in a timely manner.

I am extremely grateful to all the transcribers and coders who helped with this study. They are Ms. Alicia Christy, Ms. Audrey Risius, Mr. William Lewis, Professor Erik Voss, Professor Jim Ranalli, Ms. Erin Todey, Ms. Jenny Anderson, Professor Stephanie Link, Mr. Todd Paben, Dr. Zhi Li, Ms. Manman (Mandy) Qian, Ms. Mo

Chen, Dr. Hong Ma, Mr. Ivon Katz, Professor Dan Douglas, Ms. Felicity Douglas, Ms. Kimberly Levelle, and Professor Jesse Gleason. Many of them are my close friends. Without their help, the tedious work of spoken data transcription and coding could not have been completed.

My sincere thanks go to Ms. Teresa Smiley and Dr. Rakshak Sarda. Teresa helped straighten many things up when I was lost in filing paperwork for exams and graduation. Dr. Sarda, my cardiologist, assured me of my physical strength for pursuing my dream of being a language tester.

I am truly thankful to my colleagues and friends at the English Department and the Graduate College, including Dr. Yoo-Ree Chung, Dr. Xuan (Roger) Teng, Ms. (naughty) Deanna Ward, Mr. (naggy) Karl Schindel, Ms. (quiet) Lisa Elm, Ms. Judy Strand, Ms. Charlene Shaw, Ms. Linda Thorson, to name a few. They brightened up the days when I struggled with my dissertation.

Finally, my special thanks go to my beloved wife, Hong Wang, my parents, Guanghong Xu and Ruiying Li, and my parents-in-law, Zhian Wang and Zhiqin Xie. They firmly believed in and supported me no matter how many times I fell. When I was gloomy, they cheered me up with their unconditional love.

ABSTRACT

Collocation, known as words that commonly co-occur, is a major category of formulaic language. There is now general consensus among language researchers that collocation is essential to effective language use in real-world communication situations (Ellis, 2008; Nesselhauf, 2005; Schmitt, 2010; Wray, 2002). Although a number of contemporary speech-processing theories assume the importance of formulaic language to spontaneous speaking (Bygate, 1987; de Bot, 1992; Kormos, 2006; Levelt, 1999), none of them gives an adequate explanation of the role that collocation plays in speech communication. In the practices of L2 speaking assessment, a test taker's collocational performance is usually not separately scored mainly because human raters can only focus on a limited range of speech characteristics (Luoma, 2004).

This paper argues for the centrality of collocation evaluation to communication-oriented L2 oral assessment. Based on a logical analysis of the conceptual connections among collocation, speech-processing theories, and rubrics for oral language assessment, I formulated a new construct called Spoken Collocational Competence (SCC). In light of Skehan's (1998, 2009) trade-off hypothesis, I developed a series of measures for SCC, namely Operational Collocational Performance Measures (OCPMs), to cover three dimensions of learner collocation performance in spontaneous speaking: collocation accuracy, collocation complexity, and collocation fluency. I then investigated the empirical performance of these measures with 2344 lexical collocations extracted from sixty adult English as a second language (ESL) learners' oral assessment data collected in two distinctive contexts of language use: conversing with an interlocutor on daily-life topics (or the SPEAK exam) and giving an academic lecture

(or the TEACH exam). Multiple regression and logistic regression were performed on criterion measures of these learners' oral proficiency (i.e., human holistic scores and oral proficiency certification decisions) as a function of the OCPMs.

The study found that the participants generally achieved higher collocation accuracy and complexity in the TEACH exam than in the SPEAK exam. In addition, the OCPMs as a whole predicted the participants' oral proficiency certification status (certified or uncertified) with high accuracy (Negelkerke R^2 = .968). However, the predictive power of OCPMs for human holistic scores seemed to be higher in the SPEAK exam (adjusted R^2 = .678) than in the TEACH exam (adjusted R^2 = .573). These findings suggest that L2 learners' collocational performance in free speech deserve examiners' closer attention and that SCC may contribute to the construct of oral proficiency somewhat differently across speaking contexts. Implications for L2 speaking theory, automated speech evaluation, and teaching and learning of oral communication skills are discussed.

CHAPTER 1: INTRODUCTION

1.1 Automated Speech Evaluation (ASE): Is It a Fairy Tale?

> No operational construct definition can ever capture the richness of what happens as a process as complex as human communication, even if the speaking test is mediated by tape or computer (Fulcher, 2003, p. 19).

Fulcher's words aptly describe the extreme complexity and dynamic nature of speaking. Considering that defining the construct of oral proficiency, particularly specifying its fluidity across various contexts of language use, is a divisive issue (Bachman, 2007; Chapelle, 1998; Chapelle, Enright, & Jamieson, 2008; Douglas, 1998), it certainly sounds a fairy tale to leave the human job of assessing an individual's spoken language ability to a computer. If "[a]greement does not exist on a single best way to define language proficiency", as Chapelle et al. (2008, p. 1) asserted, then based on what gold standards can a computer be programed to evaluate human language? Bearing in mind this conundrum, readers may be very surprised to learn that fully automated oral English assessments, such as Versant English Test and Test of English as a Foreign Language (TOEFL) iBT Speaking Practice Test, have been in the market for years.

Regarding the credibility and quality of the aforementioned automated oral language assessments, it is unlikely that the following issues are unquestioned. First, are we there yet? In other words, is the state-of-the-art technology advanced enough to process and evaluate complex human speech communication? Second, what speech elements does the computer actually score? Third, to what extent can we trust the claims made by test developers about the automated

scores? The first question concerns the technological constraints on automated speech evaluation (ASE), the second question the construct coverage and relevance of ASE, and the last question the validity of the interpretation and use of ASE results. All these questions and doubts point us in the direction of the much-needed research on ASE and inspired this dissertation study. The next section reviews two major types of ASE being used today and discusses why one is more promising than the other.

1.2 Background of the Study

As of today, ASE generally falls in two types. The two types can also be viewed as two main approaches to realizing fully automated speaking assessment. The first type, represented by Pearson's Versant speaking tests (formerly the PhonePass Test, Chun 2006), assesses highly restricted oral responses, including reading sentences aloud, repeating sentences aloud, saying opposite words, giving short answer to questions, building a sentence from phrases, answering opinion questions, and retelling spoken passages (Bernstein, Van Moere, & Cheng, 2010, p. 358). I call this type of speaking tests restricted ASE or RASE because test takers' speaking performance is elicited by highly controlled language production tasks. The second type, such as Education Testing Service's (ETS's) TOEFL iBT Speaking Practice Test, assesses extended, open-ended oral responses such as speaking on familiar topics, integrated reading and speaking, and integrated listening and speaking (Xi, 2008). Although speech length of such tasks is also controlled at this point, the automated scoring program that supports the examination has the potential of evaluating speech of free will and infinite length. Accordingly, I call this type of automated speaking tests unrestricted ASE or UASE. There are other less prominent automated speaking assessments in the market, such as Carnegie Speech Assessment (Carnegie Speech,

2015) and SpeechTrac (MeritTrac, 2015), that may fall in the second category. However, there is relatively scarce information about their test structures and the intended constructs underlying the assessments. For this reason, I choose to use Versant and TOEFL iBT Speaking Practice Test as examples below.

It is generally assumed that the construct of speaking comprises high-level cognitive processes that are responsible for formulating and lexicalizing concepts and low-level production/articulation skills that are responsible for turning the mentally planned internal speech into overt speech; in addition, the process of proficient speech formulation involves close coordination of the various processes and skills so that the speaker's limited attentional resource can be spent most economically under the time pressure (e.g., Bygate, 1987; de Bot, 1992; Kormos, 2006; Levelt, 1999a, 1999b). In the following sections, I will argue that the construct of RASE is limited in that it mainly assesses a speaker's low-level production skills; it is UASE that has the potential of assessing the construct of speech formulation as a whole (see Figure 1.1).

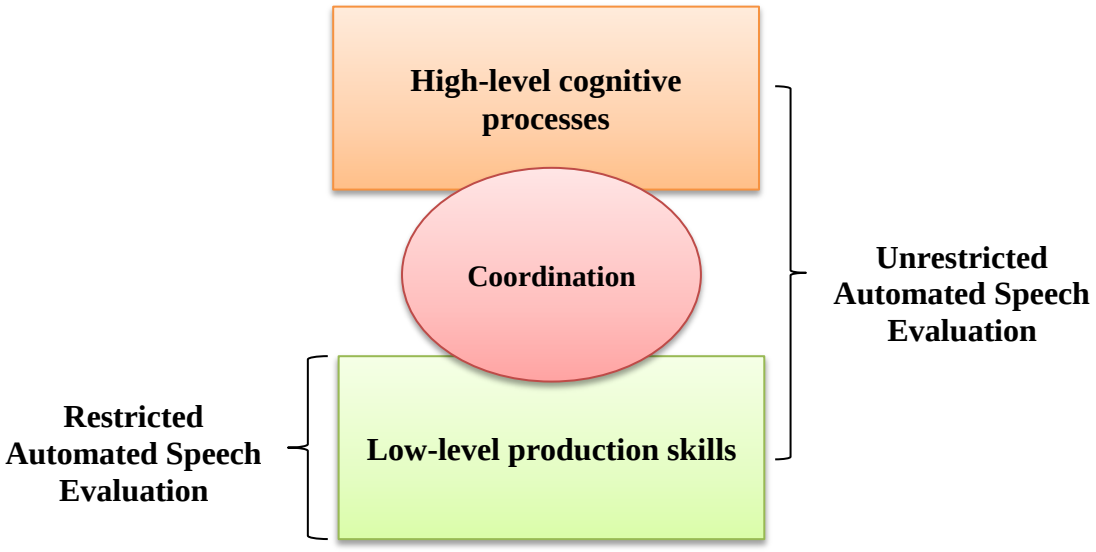


Figure 1.1. A simple illustration of the construct coverage of UASE and RASE

1.2.1 Restricted Automated Speech Evaluation (RASE)

RASE features good accuracy of voice-to-text recognition but a low degree of authenticity. As constrained oral responses are short and highly predictable, automated recognition of non-native, accented speech based on matching appears to be perfectly feasible and reliable. Nonetheless, RASE is criticized for eliciting minimal and decontextualized language production, thus significantly altering the nature of speaking (see e.g., Chun, 2006, 2008).

Arguably, the language constructs that RASE engages is markedly different from what real-world oral communication normally does. First, RASE chooses not to elicit a nonnative speaker's discourse-level patterns of language use which, however, has been found to affect native-speaking listeners' perception of the coherence and comprehensibility of speech (Rossiter, Derwing, Manimtim, & Thomson, 2010; Tyler, 1992).

Second, RASE eliminates some important underlying processes of speech generation from the test construct including analyzing language use situations (e.g., formality of the situation), conceptual generation (i.e., planning on the message that the speech communicates), lemma selection (i.e., choosing words), syntactic composition (i.e., forming sentence structures), applying compensatory strategies (i.e. anticipating or solving problems in speech formulation), and so forth (Bygate, 1987; Levelt, 1999a, 1999b). These processes, as will be discussed in more detail in Chapter 2, Section 2.3, are believed to contribute significantly to the quality of speech product in real-world communicative situations (see more discussion in Chapter 2). In other words, it seems that RASE only assesses the "superficial layer of English speaking ability" instead of its core components (Fan, 2014, p. 12).

Third, as RASE mainly elicits context-reduced language performance, it fails to engage the contextual components in the oral construct such as strategic competence and sociolinguistic competence. Research on language for specific purposes (LSP) testing has suggested that language proficiency is not an absolute, context-independent conception; rather, it is a fluid conception the meaning of which is largely determined by the socio-political context of language use (e.g., Douglas, 2000; Douglas & Selinker, 1992; Halleck & Moder, 1995; Moder & Halleck, 2009, 2012). More recently, language testers seem to be more inclined to take an interactionalist perspective on construct definition which attributes the observed language performance to the interactions between learner attributes and situational factors (see e.g., Bachman, 2007; Bachman & Palmer, 2010; Chapelle, 1998; Douglas, 2000). The latest argument-based validation framework, which "is proving useful in language assessment" (Chapelle, 2012a, p. 26), also begins with defining the typical acts of communication in the target language use (TLU) domain (see e.g., Chapelle, et al., 2008; see also the notions of 'situational authenticity' and 'interfactional authenticity' in Bachman, 1991).

Lack of authenticity turns out to be an insurmountable obstacle in validating the interpretation and use of RASE scores if the test developers claim that the test contents are authentic. For example, Downey, Farhady, Present-Thomas, Suzuki, and Van Moere (2008) argue that the Versant test tasks "do demonstrate—at the very least—an appreciable degree of authenticity" (p. 164). However, the constrained speaking tasks, as discussed above, at the first place, alter the nature of speaking and create a noticeable gap between the test construct and the L2 oral construct in well-accepted theory (see more discussion below). In this sense, authenticity not only concerns the correspondence between the characteristics of test tasks and those of TLU tasks (Bachman & Palmer, 1996, p. 23) but also construct relevance and representation, i.e., the

degree to which the construct underlying test performance overlaps with that underlying the TLU performance (Messick, 1989, p. 39).

The test construct of RASE is too narrow to account for real-world speaking behaviors. The Versant speaking test, for example, purports to assess a 'facility construct' that is defined as "the ability to understand the spoken language on everyday topics and to speak appropriately in response at a native-like conversational pace in an intelligible form of the language" (Bernstein, et al., 2010, p. 358). Based on this construct definition, the test developers claimed that the Versant test scores reflect some core spoken language skills that are "used in all language situations" and "underlie the ability to perform all of the communicative functions" (p. 373). These skills include understanding the meanings of the words and phrases in incoming speech, producing appropriate and prompt (short) spoken responses, maintaining native-like oral fluency, and articulating speech with accurate pronunciation (p. 362).

However, in view of the intricacy of speech processing in theory (see Chapter 2 for more detail), it seems highly problematic to extract so-called 'core skills' from the oral construct. On the one hand, there is no ground for assessing the oral construct as separate components (i.e., via a psycholinguistic approach), considering that the crux of L2 speaking is making effective use of the limited attentional resources to coordinate multiple mental processes simultaneously under the time pressure. It thus can be argued that the oral construct must be assessed as a whole (i.e., via a communicative approach) in order for L2 learners to perform this extremely cognitively demanding coordination. On the other hand, it is rather presumptuous of Versant tests to generalize about these 'core spoken language skills' given our limited knowledge of the oral construct. As a matter of fact, there is considerable evidence against the invariance of the oral construct in the real world (see e.g., Chalhoub-Deville, 1995; Schmidgall, 2013). For example,

highly accurate and intelligible pronunciation is not seen in fast-paced auctioneers' language (Kuiper, 1996); speech accentedness may have a larger effect on comprehensibility among native-speaking listeners than non-native-speaking listeners (Jun & Li, 2010). It thus seems reasonable to believe that the core components of the oral construct may vary in different contexts of language use.

Although the machine-generated Versant test scores exhibited high correlations with human criterion scores of oral proficiency (Bernstein, et al., 2010; Downey, et al., 2008), no plausible construct theory would help explain this relationship. Bernstein et al.'s (2010) theory that one's communicative spoken skills can be manifested in "activities that are not communicative" (p. 358) does not sound reasonable. It is obvious that speakers allocate attentional resources differently in communicative and non-communicative tasks (e.g., free speech vs. reading a script). Validity, according to the latest Standards for Educational and Psychological Testing, refers to "the degree to which evidence *and* [emphasis added] theory support the interpretations of test scores for proposed uses of tests" (AERA, APA, & NCME, 2014, p. 11). Without the support of a plausible construct theory, neither construct-based score interpretation nor extending the meaning of the test scores beyond the test domain would be justifiable (see Kane, 2009).

Construct-based score interpretation is critical to non-communicative language tests such as Versant in which the characteristics of the test tasks and test behaviors differ substantially from TLU tasks and behaviors. The very logic that can support the extrapolation from the observed test performance to other test domains or to the TLU domains is that the seemingly varied language behaviors can be attributed to the same construct or belong to the same 'class' (Messick, 1989, p. 41). It is creating a construct or a logic class that allows us to make inferences

about the unobserved from the observed (Russell, 1915). If an interactionalist perspective is taken, then the construct or logical class of a language test can be viewed as some hypothesized consistencies that reside both in the language learner and the context of language use (Chapelle, 1998).

As above, the ambitious claim made by Versant tests that the automated scores "may be useful across many target domains as part of a battery of tests for specific-domain speaking skills" (Bernstein, et al., 2010, p. 357) is questionable. Fan's (2014) research, for example, suggests that Versant examinees actually do not have sufficient confidence in interpreting the automated scores in this way. However, the Versant test developers did caution against using the test alone for high-stakes decision-making and suggested using the test only for a first-step screening purpose to save human testing resources (Bernstein, et al., 2010, p. 372).

In sum, RASE may make automated scoring of oral responses perfectly feasible. However, the target construct of RASE deviate from test takers and researchers' common perception of L2 oral proficiency that pertains to natural speaking behaviors. As a result, RASE has often been criticized for swapping authenticity and validity for convenience and practicality (Chun, 2008, p. 170).

1.2.2 Unrestricted Automated Speech Evaluation (UASE)

UASE, meanwhile, imposes little constraint on the way that language performance is elicited. Ideally, UASE is able to score and provide instant feedback to free oral production in task-based, communicatively oriented language assessments that simulate the TLU environment.

The TOEFL iBT Speaking Practice test, for example, uses an automated scoring program, called SpeechRater, to score naturally occurring L2 spoken responses. This low-stakes

practice test mirrors the content and design of the operational TOEFL iBT Speaking Section that measures "the academic English-speaking abilities of nonnative speakers who plan to study at English-medium institutions for higher education" (Xi, Higgins, Zechner, & Williamson, 2008, pp. 5-6). According to the test developers, the automated score is "a prediction of the score on the TOEFL iBT Speaking Practice test a test taker would have obtained from trained human raters"; it is suggested that the score be used by the test takers to "self-evaluate their readiness to take the TOEFL iBT Speaking test" (Xi, 2008, pp. 109-110). The current version of SpeechRater only produces a holistic score without explaining the rationales behind scoring. However, it is a long-term goal of the TOEFL testing program to equip SpeechRater with the ability to provide immediate instructional and diagnostic feedback in addition to a single score (Xi, 2008).

Unfortunately, there is no such thing as a free lunch. The extended spoken responses create enormous difficulty for automated speech recognition. At present, low accuracy in speech-to-text recognition seriously restricts the construct coverage of UASE. SpeechRater, for example, is only capable of scoring a subset of speech features that do not require accurate speech recognition, such as speech rate, duration of silences, average number of pauses, and the ranges of pitch contours (see Chapter 2, Section 2.5.2 for a more detailed review). As the scoring criteria are not strictly meaning-based, test developers are concerned that test takers may trick the scoring system by utilizing inappropriate response strategies (Xi, 2010). That is, if the scoring algorithms "fail to assign credit to qualities of a response that are relevant to the construct that the test is intended to measure" (Chapelle, & Douglas, 2006, p. 41), test takers may choose to ignore these qualities in language production. These gaming behaviors would not only undermine the validity of construct-based score interpretation but also lead to negative consequences of testing—the way high-stakes language tests are designed has a huge impact on

the societal perception of language proficiency and the way language is learned and taught at schools (Xi, 2012).

Even though UASE scores were found to correlate strongly with human criterion scores (Xi, et al., 2008), test developers were keenly aware of the big discrepancies between the scoring algorithms and human rating criteria (Xi, 2008, 2010). SpeechRater, according to Xi (2010), is not ready for high-stakes testing purposes although such a need is paramount.

1.2.3 The Great Promise of UASE

Despite its present drawback in construct coverage, UASE undoubtedly is more promising than RASE. As stated by the latest *Standards for Educational and Psychological Testing*, "[w]hen automated algorithms are to be used to score complex examinee responses, *characteristics of responses* [emphasis added] at each score level should be documented along with the theoretical and empirical bases for the use of the algorithms" (AERA, APA, & NCME, 2014, p. 91). It would be hard for RASE to meet this requirement because its scoring model is fundamentally based on matching rather than a careful analysis of the construct-relevant characteristics in learner language (see Bernstein, et al., 2010, p. 361 for an example). Only UASE has the potential for fine-grained content analysis of learner language based on which detailed descriptors about the language forms and features at each score level can be summarized and relevant diagnostic feedback can be generated.

Recently, research on automated speech-to-text recognition has made important breakthroughs (see Rashid, 2012; Yu & Deng, 2014). The accuracy of computerized transcription of foreign accented English speech has increased from 51.4% (Xi, Higgins, Zechner, & Williamson, 2012) to 72.0% (X. Wang, Evanini, & Zechner, 2013). It is foreseeable

that a computer will eventually be able to transcribe learner speech at near-perfect accuracy in the near future. When reliable L2 speech recognition technologies were easily obtainable, an acute problem that UASE developers would face is to build interpretable automated scoring rules to adequately and accurately represent the fluid concept of L2 oral proficiency. The development of such scoring rules, as discussed below, is a critical piece in evaluating the validity of the interpretation and use of automated scores.

1.3 Statement of the Problem

The problem that this dissertation study tried to solve arose from the need to expand the construct coverage of UASE. To clearly address this research need, I frame the problem in a hypothetical interpretive/use argument (IUA) for using automated score to resolve the score discrepancies between two human raters and using automated test feedback to facilitate language learning and instruction in a university-based oral English assessment for international teaching assistants (ITAs). By hypothetical, I mean that this aim of using automated scores to complement human scores, at this point, is not pursued by any ITA testing program in the real world. However, I recognize the implementation of automated scoring as a burning need to a future generation of oral English language assessments that are expected to produce test results efficiently and test feedback in great detail (see a further discussion in Chapter 2, Section 2.5). In fact, many researchers share my vision and have invested considerable research efforts in this direction (e.g., Enright & Qian, 2010; Weigle, 2010; Williamson, Xi, & Breyer, 2012; Xi, 2010).

1.3.1 What is an Interpretive/Use Argument?

Constructing an IUA is the first step in performing argument-based validation. An IUA consists of "a network of inferences and assumptions inherent in the proposed [test score] interpretation and use" (Kane, 2013, p. 2); it serves as a "conceptual tool" for unfolding "the multifaceted meaning of test scores" (Chapelle, 2012b, p. 19), evaluating and prioritize validation research needs, and organizing various sources of validity evidence obtained in a coherent manner. An IUA can be tailored for particular testing purposes. That is, depending on the intended score use, the IUA may comprise different elements (Kane, 2009).

An IUA or an interpretive argument provides the structure of a validity argument. That is, it can be made into, at best, a plausible argument for the proposed score interpretation and use if all of the inferences and assumptions it entails are supported to the extent possible and all of the challenges (e.g., alternative interpretations) that pose potential threats to these inferences and assumptions are weakened or disproved with reasoning or evidence (Kane, 2013).

1.3.2 A Hypothetical Interpretive/Use Argument for a Semi-Automated ITA Oral English Assessment

Let us imagine a scenario in which a proposal is made to use an automated scoring program in an ITA Oral English Assessment—a test used to determine ITAs' readiness for teaching at an English-medium university—in order to make the score reporting more efficient and informative. Specifically, it is proposed that an automated score is only used to resolve score discrepancies between two human raters beyond a certain threshold; that is, the automated score and human scores are averaged to produce the reported test score. It is also proposed that the diagnostic feedback generated by the automated scoring program based on its scoring features

(i.e., the speech features it is trained to evaluate) be provided to the examinees and their ESL instructors along with the reported test score. It is hoped that this feedback is useful for self-learning and teaching of the oral communication skills necessary for performing university teaching duties.

The hypothetical IUA displayed in Figure 1.2 is created to direct validation research that aims to support the proposed use of this imagined automated scoring program in the ITA oral English assessment. It is noteworthy that this particular IUA does not serve as the skeleton of the primary validity argument for the ITA Oral English Assessment which should be more comprehensive in scope and cover issues on item development, test administration, human scoring, and so forth—it is not my aim in this paper to enter into the details of the primary validity argument (see Chapelle, et al., 2010 for a good example). However, the hypothetical IUA can be viewed as a supplement to the primary validity argument that addresses additional issues arising from the proposed use of automated scoring.

The hypothetical IUA contains four inferences: evaluation, explanation, utilization, and ramification. To make the IUA plausible, each inference has to be authorized by a warrant that rests on the supports from some underlying assumptions. Put simply, the chain of inferences in the IUA can be viewed as a sequence of bridges that a test user must cross to confidently accept the proposed use of the automated scoring program in the ITA oral English assessment. A warrant is analogous to a ticket that authorizes the crossing of an inferential bridge. The assumptions are like the conditions that must be met for issuing the ticket.

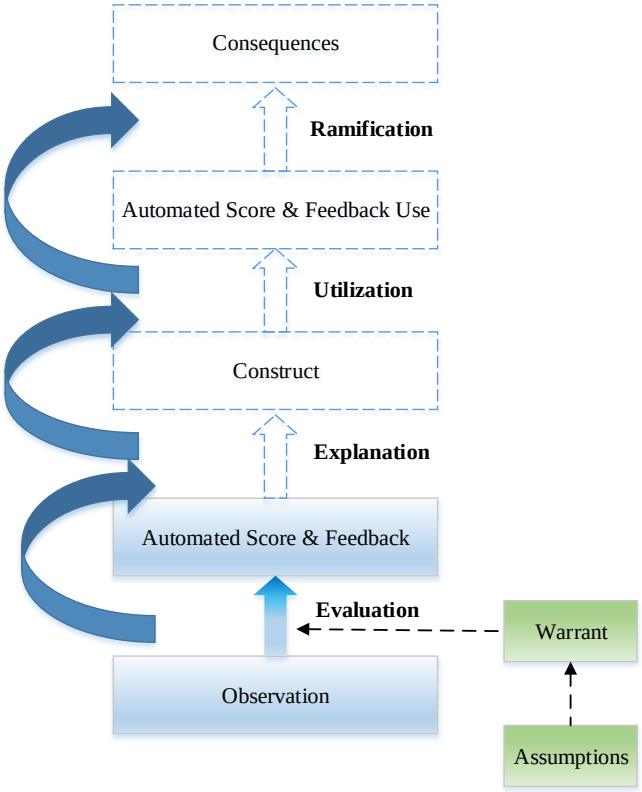


Figure 1.2. The interpretive/use argument for using automated score for adjudication in the case of large discrepancies between two human ratings

The evaluation inference at the bottom of the IUA directly pertains to automated scoring. The crossing of this inference is authorized by the warrant that an examinee's observed test performance (i.e., the recorded speech samples) is evaluated appropriately to yield an automated score and diagnostic feedback reflective of the construct of academic oral proficiency. There are four pivotal assumptions underlying this warrant. First, construct-relevant speech features (i.e., observed speech qualities that contribute to academic oral English proficiency) can be precisely defined. Second, the computer algorithms can accurately and comprehensively capture these features in the L2 spoken data. Third, the computer algorithms can evaluate these features in a way that reflect the quality of language use. Fourth, the way the automated features are weighted and combined to produce a holistic automated score reflects the importance of each feature to the

targeted construct of academic oral proficiency according to theory or relevant research. A potential rebuttal to this inference is that the automated scoring program fails to assess some key components in the targeted L2 oral construct (owing to technological constraints).

The ensuing explanation inference concerns the construct-based interpretation of the automated score. The crossing of this inference is licensed by the warrant that the automated score is attributed to the targeted construct of academic oral proficiency. This warrant rests on two assumptions. First, the use of automated score does not change the meaning of the reported test score (Xi, 2010). Second, the automated feedback is relevant to the examinee performance and the construct of academic oral proficiency. If the rebuttal from the evaluation inference mentioned above is not offset, a potential rebuttal that the explanation inference may confront is that the constructs embodied by the human score and automated score are obviously disparate. In other words, the meanings of an automated score and a human score are intrinsically different.

The utilization inference that follows evaluates the usefulness of the automated score and feedback. It is authorized by the warrant that the automated score is useful for resolving human score discrepancies and the automated feedback is useful for identifying the learning needs of the examinees. The warrant is supported by three assumptions. First, the automated score accurately predicts a test score that could have been assigned by a trained human rater. Second, the automated diagnostic feedback is clearly interpretable by examinees and their ESL teachers. Third, the automated diagnostic feedback is useful for making decisions on examinees that fall on the cut-off scores and for placing examinees into appropriate remedial classes. As a consequence of the rebuttal from the previous explanation inference, a potential rebuttal against the utilization inference would be against the policy of combining automated and human scores

to produce a reported test score. That is, there is no logical ground for combining two scores that obviously convey different meanings.

The last ramification inference appraises the consequences of the proposed use of automated scoring in the ITA oral English assessment. The crossing of this inference is authorized by the warrant that the proposed use of the automated scoring program generally has a positive impact on the stakeholders of the testing program including examiners, examinees, ESL teachers, and the students being taught by the examinees. This warrant is buttressed by four assumptions. First, test users understand the automated scoring procedure and this understanding has a positive impact on teaching and learning of the oral English communication skills necessary for teaching. Second, examinees and their ESL teachers benefit from receiving the automated diagnostic feedback. Third, the use of automated score reduces the cost of testing and the time of score reporting that would otherwise be spent on hiring an additional human rater. Fourth, the use of automated score does not have a negative impact on the test-taking strategies that examinees employ (e.g., gaming behaviors mentioned previously). If the rebuttal to the earlier utilization inference is not fully discounted, the ramification inference likewise will be under attack by a potential rebuttal claiming that the use of automated score would lead to criticism, skepticism, resistance, or even hostility to the ITA oral English assessment. A summary of the warrants, assumptions, and potential rebuttals associated with the four inferences are presented in Table 1.1 below.

Table 1.1 A Summary of the Warrants, Assumptions, and Potential Rebuttals Associated with the Evaluation, Explanation, Utilization, and Ramification Inferences (adapted from Chapelle, et al., 2008, 2010)

Inference	Warrant Licensing the Inference	Assumptions Underlying the Warrant	Potential Rebuttals
Evaluation	An examinee's observed test performance is evaluated appropriately to yield an automated score and diagnostic feedback reflective of the construct of academic oral proficiency.	1. Construct-relevant speech features (i.e., observed speech qualities that contribute to academic oral proficiency) can be precisely defined. 2. The computer algorithms can accurately and comprehensively capture these features in the L2 spoken data. 3. The computer algorithms can evaluate these features in a way that reflect the quality of language use. 4. The way the automated features are weighted and combined to produce a holistic score reflects the importance of each feature to the construct of academic oral proficiency according to theory or relevant research.	1. The automated scoring program fails to assess some key sub-components in the targeted L2 oral construct (e.g., collocation)
Explanation	The automated score is attributed to the targeted construct of academic oral proficiency.	1. The use of automated score does not change the meaning of the reported test score. 2. The automated feedback is relevant to the examinee performance and the construct of academic oral proficiency.	1. The constructs embodied by the human score and automated score are obviously disparate.
Utilization	The automated score is useful for resolving human score discrepancies and the automated feedback is useful for identifying examinees' learning needs.	1. The automated score accurately predicts an expert human score that could have been used. 2. The automated diagnostic feedback is clearly interpretable by examinees and their ESL teachers. 3. The automated diagnostic feedback is useful for making decisions on examinees that fall on the cut-off scores and for placing examinees into appropriate remedial classes.	1. It is unreasonable to combine two scores that convey different meanings.

Table 1.1 (continued)

Inference	Warrant Licensing the Inference	Assumptions Underlying the Warrant	Potential Rebuttals
Ramification	The proposed use of the automated scoring program generally has a positive impact on the stakeholders including the examiners, the examinees, the ESL teachers, and the students being taught by the examinees.	1. Test users understand the automated scoring procedure and this understanding has a positive impact on teaching and learning of oral English communication skills. 2. Examinees and their ESL teachers benefit from receiving the automated diagnostic feedback. 3. The use of automated score reduces the cost of testing and the time of score reporting that would otherwise be spent on hiring an additional human rater. 4. The use of automated score does not have a negative impact on the test-taking strategies that examinees employ.	1. The proposed use of automated score would lead to criticism, skepticism, resistance, or hostility to the ITA Oral English assessment program.

The above hypothetical IUA has shown that validity concerns on automated scoring do not stay confined within the evaluation inference; instead, they permeate through an IUA (Bennett, 2004; Bennett & Bejar, 1998; Williamson, Bejar, & Mislevy, 2006; Xi, 2010). However, an evaluation inference concerning the design of the automated scoring rules is the weakest link in an IUA for UASE because most serious challenges may arise from it. That is, a strong rebuttal to the evaluation inference (e.g., a critique on the construct underrepresentation of an automated scoring program), as shown above, can cause a 'ripple effect' on the entire IUA, weakening the following inferences (on score interpretation, score use, and the consequences of this interpretation and use) one after another (diagramed as curved arrows in Figure 1.1). As this weakest link limits the overall plausibility of the inferential chain, it deserves the most attention from test developers and validators (Kane, 2013, p. 15). Probably for this reason, many

researchers anticipate that the validation efforts on UASE will likely concentrate on the interactions among automated scoring (evaluation), score meaning (explanation), and test consequences (ramification) in the following decades (e.g., Xi, 2010; Williamson, et al., 2012).

1.3.3 Collocation and Automated Speech Evaluation

Construct underrepresentation of UASE, as discussed in the previous sections, is a major obstacle to using automated scores for high-stakes decision-making. The impetus for this dissertation study came from the urgent need to expand the construct coverage of automated scoring for communicative oral language assessment (Xi, 2010). However, this construct expansion is not simply a technological problem. Asides from technological feasibilities, UASE developers need to seriously consider what speech features are construct-relevant and how to set up a scoring model based on these features. Such questions solicit fundamental research on the meaning of the construct of L2 oral proficiency or a consistent theory about how this hypothetical entity can be elucidated in terms of observed variables in specific TLU domains (see a further discussion in Chapter 2, Section 2.5.3).

Collocations are fundamental building blocks of language. These formulaic expressions have been found to pose serious problems for L2 learners (e.g., Voss, 2012). Many researchers argue that effective use of collocations in L2 speaking is a defining aspect of L2 oral proficiency (e.g., Handl, 2008; Wood, 2010). Given the importance of collocation to L2 speaking, it seems almost inexcusable for UASE to ignore the observed collocation occurrences in spontaneous L2 speech.

Then, the question is how the observed collocation occurrences can be measured to generate scores not only useful for predicting human judgment of L2 oral proficiency but also

interpretable to laypersons? Stated another way, what observed qualities in the collocations extracted from learner speech are indicative of oral proficiency? As will be discussed in Chapter 2, Section 2.5, speech features that adequately predict human criterion scores are not necessarily logically interpretable. For developing theoretically sound and empirically useful collocation measures, it is clear that the theories of speech processing (i.e., the process of speaking), the conception of second language (L2) oral proficiency (i.e., evaluation criteria of the construct), and collocation must be consulted and conceptually connected.

Besides, although theory asserts a relationship between spoken collocation usage and perceived oral language proficiency, previous literature has provided very limited empirical knowledge of its magnitude. In the future, UASE developers will likely need this information to specify the automated scoring model. The current UASE scoring programs such as SpeechRater rely on experts' opinions to determine weight assignment on each scoring feature in the scoring model (Xi, et al., 2012). However, "experts are not infallible" (Clauser, Kane, & Swanson, 2002, p. 426; see also Chapelle & Chung, 2010, p. 309). The subjective expert-articulated scoring criteria require empirical justification, too, to discount prospective challenges and doubts against their validity.

Finally, a foreseeable challenge that UASE developers will confront is to determine whether and how to tweak the scoring algorithms for responses elicited in different settings of language use? This is an issue germane to the construct definition of oral language proficiency. That is, is the targeted oral construct context-invariant or is there a somewhat different conception of language ability accounting for each situation of language use covered by the test? I believe many language testers would vote for the latter but have slightly different opinions on the degree of the construct-context interaction (see e.g., Bachman, 2007; Chalhoub-Deville,

1995, 2003; Chapelle, 1998; Douglas, 2000). If the targeted oral construct is fluid, then there must be a collocation-context interaction as well, meaning that collocation usage may contribute to perceived oral proficiency at least differently in various speaking contexts. Yet this hypothesis needs to be backed up with empirical evidence (backing for assumption 4 mentioned above).

1.3.4 Inspiration from Previous Research

Xu and Xi (2010) had some preliminary but promising findings on the relationship between non-native English speakers' collocation usage in spontaneous speech and perceived oral English proficiency. Based on 556 speech samples collected from the TOEFL iBT Speaking Practice Test, they found that the normalized frequency of the collocational errors in learner speech had a weak correlation (r = -.196, p < .01) with the human criterion scores of oral proficiency; interestingly, the magnitude of this relationship became stronger (r = -.299, p < .01) when the speech sample became longer (i.e., when the responses produced by the same speaker across test tasks were aggregated).

Xu and Xi's study is limited in three aspects. First, low-level English speakers were underrepresented (4%) in their sample. With a negatively skewed score distribution, the estimated correlation coefficient could be attenuated. Second, their speech samples were relatively short (only 45-60 seconds in length), thus containing insufficient number of collocation occurrences. For reliable measurement of one's spoken collocational competence (i.e., the ability to make effective use of collocations to enhance speech performance; see Chapter 2, Section 2.4 for a precise definition), it seems that longer speech must be elicited. Third and most importantly, Xu and Xi's collocation measure was unidimentional, only taking into account the accuracy/acceptability of lexical combination. However, a few second language

acquisition (SLA) researchers have pointed out that language learners at different proficiency levels may use the target language with different goals, sometimes focusing on accuracy and fluency whereas other times on sophistication or complexity (e.g., Skehan, 1998; Robinson, 2001; Ellis & Barkhuizen, 2005). Xu and Xi, for example, noticed that low-level speakers were inclined to use simple collocations possibly with the intention to lessen the chance of making mistakes. Therefore, multidimensional collocation measures that can fully capture the characteristics of learner collocation usage in speaking need to be developed.

As above, for a finer-grained understanding of the relationship between English as a second language (ESL) learners' actual collocation use in speaking and their perceived oral English proficiency and how this relationship interacts with the contexts of language use, a study that elicits adequately long learner speech in distinct contexts of language use, that develops sophisticated collocation measures, and that selects a balanced sample of ESL learners spreading over varied oral proficiency levels is needed.

1.4 Goals and Significance of the Study

This research study serves as a foundational research project for automated speech evaluation. Its primary goal is to investigate the empirical performance of a series of theory-based collocation measures on learner spoken data for predicting human criterion scores of oral English proficiency. The finding will directly inform the development of the collocation evaluation component in SpeechRater and other similar UASE systems that are being developed.

A second goal of the study is to test the hypothesis that the relationship between spoken collocational performance and perceived oral English proficiency is affected by the speaking context. This finding will provide evidence for language testers' ongoing debate on whether

contextual factors belong to the intended test construct. According to Messick (1980, p. 1019), "At any point new evidence may dictate a change in construct, theory, or measurement." Thus, if the above hypothesis is supported by this study, we then have more confidence in believing in the claim that "the nature of the L2 oral construct is not constant" (Chalhoub-Deville, 1995, p. 251).

A third goal of the study is to draw language testers' attention once again to construct validation. Kane's argument-based framework may help "remov[e] the enormous burden that otherwise be placed on an imprecise theoretical construct" in test validation (Chapelle, Enright, & Jamieson, 2010, p. 12). Nevertheless, this does not seem to be the case for validating an automated language test. Since computer algorithms always rely on very specific instructions to evaluate language data, it seems impossible to circumvent a fine-grained construct definition of language proficiency. That is, until we are able to clearly articulate what a language construct means in terms of observed language performance characteristics and variables, we will not have enough confidence in building a trustworthy automated scoring model and explaining the meanings of automated scores to stakeholders.

1.5 Outline of Dissertation

The ensuing Chapter 2 reviews the literature on the construct definition of collocation, the relationship between collocation and L2 learning, and theoretical models of speech formulation. The literature review leads to a proposal of a new construct: spoken collocational competence (SCC). It is argued that SCC is distinctive from collocational knowledge that cued collocation tests engage and that measuring SCC is the key to measuring L2 oral proficiency. Chapter 3 presents the methodology of this study. In this chapter, the development of the

operational SCC measures and the procedure of validating these measures are detailed. Chapters 4 and 5 report the results of the study and discuss the implications of the findings, respectively.

CHAPTER 2: LITERATURE REVIEW

Both ‘collocation’ and ‘oral language proficiency’ are hypothetical constructs conceived by language researchers to account for the consistencies in observed human language behaviors. An investigation into the relationship between the two hence must be preceded by a clear understanding of their meanings and the ‘nomological net’ (i.e., the underlying theory) that logically link them together (Cronbach & Meehl, 1955). This chapter consists of eight sections. The first two sections review the various perspectives taken in defining the construct of collocation and the large body of empirical research regarding the relationship between collocational knowledge and L2 proficiency. The third and fourth sections tap into speech-processing models and spoken collocational competence (SCC), a new construct derived from a logical analysis of the role that collocation plays in speech formulation. The fifth section gives an overview of SpeechRater’s design and the major obstacles it currently faces. The sixth section reviews the cutting-edge technology of automatic collocation extraction. The chapter ends with a brief summary and a statement of the specific research questions of this study.

2.1 Defining the Construct of Collocation

Since Palmer (1933) introduced the concept of collocation nearly eighty years ago, there unfortunately has never been a consistent way of defining the term (Handl, 2008; Nesselhauf, 2005; Read & Nation, 2004; Schmitt, 2010; Wray, 2002). The conception of collocation is elusive mainly because language researchers from a wide range of disciplines, such as phraseology, lexicography, corpus linguistics, second language acquisition, and psycholinguistics, emphasized different defining criteria of this language phenomenon. These criteria include, for example, the statistical tendency of word co-occurrence (Biber, Conrad, &

Leech, 2002), holistic storage of words in the linguistic system (Conklin & Schmitt, 2008; Ellis, 1996; Pawley & Syder, 1983; Sinclair, 1991), the semantic prosody between words (Carter, 1998; Firth, 1968; Kennedy, 2008), the degree of freedom for lexical substitution (Benson, Benson, & Ilson, 1986; Cowie, 1978; Fernando, 1996; Nesselhauf, 2005) and a multi-dimensional combination of the above-mentioned criteria (Handl, 2008).

Schmitt (2010, p. 119) made an analogy to describe the various approaches taken to identifying collocation and formulaic language:

> Just as the five blind men of Hindustan who went out to learn about an elephant felt different parts of the elephant's body and came to very different conclusions about what an elephant is like, researchers seem to be looking at different aspects of formulaic language and using terminology to make sense of that aspect.

By combining the diverse views on collocation, I was able to summarize four important aspects of the concept. First, collocation is the co-occurrence of two or more words but it mainly refers to two-word pairs. This, according to many researchers, is fundamental for discussing collocation (e.g., Biber, et al., 2002; Handl, 2008; Nesselhauf, 2005; Schmitt, 2010; Sinclair, 1991). Co-occurrence, however, does not mean that two words must be next to each other. Rather, the parts of a collocation may or may not be separated (Halliday & Hasan, 1976). Collocation researchers generally believe that such co-occurrence does not result from random choices of words but is constrained by certain relations between the words, such as topic, register, style, sociolects, semantic prosody, connotation, lexical repulsion, and grammaticality (Bednarek, 2008; Renouf & Banerjee, 2007, 2008; Schmitt & Carter, 2004; Whitsitt, 2005).

Among these constraints, semantic prosody, which determines the meaningfulness of combination, is arguably the most important (e.g., Barnbrook, 2007; Moon, 2008; Sinclair, 1991).

Second, collocation is often divided into two categories, i.e., grammatical collocation and lexical collocation (e.g., Bahns, 1993; Benson, et al., 1986; Cowie, 1981; Durrant, 2009; Gries, 2008; Schmitt, 2000). A grammatical collocation consists of a dominant content word (a noun, an adjective, or a verb) and a subordinate grammatical structure (such as a preposition, an infinitive, a gerund, or a clause). For example, phrases like *interested in*, *adhere to*, *at night*, and *in advance* are grammatical collocations. In contrast, a lexical collocation is combined by two content words that contribute almost equally to its whole meaning. Some examples of lexical collocations are *throw-party*, *bee-buzz*, *movie-theater*, and *heavy-smoker*. It is noteworthy that lexical collocations can be consistently classified according to their syntactic patterns. For instance, *throw a party* is a verb-noun collocation and *a heavy smoker* is an adjective-noun collocation. This syntactic approach was taken in identifying collocations in this study (see further discussion in Chapter 3).

Third, collocation is deemed by phraseologists as a stretch of formulaic language somewhere in the middle of the continuum of idiomaticity (Figure 2.1), with fixed expressions on one end and free combinations on the other (Benson, et al., 1986; Carter, 1998; Cowie, 1978, 1981; Ellis, 2008; Fernando, 1996; Howarth, 1998; Nesselhauf, 2005). In determining where a phrase falls on the continuum, two gradable criteria are examined: substitutability (i.e., to what degree can its elements be freely substituted) and transparency (i.e., to what degree is the meaning of the expression literal?). Fixed expressions such as idioms, proverbs, and compounds are the least open to substitution and the most opaque in meaning. It is believed that fixed

expressions are holistic units that can be directly accessed in memory without full linguistic analysis (Glucksberg, 1993; McGlone, Glucksberg, & Cacciari, 1994). Take *a piece of cake* (easy to do) as an example. Its figurative meaning is lost if you say 'the piece of cake', 'a slice of cake' or 'a piece of pancake'. Free combinations (also known as free collocations, open collocations, or casual collocations), on the other hand, "consist of elements used in their literal senses and freely substitutable" (Howarth, 1998, p. 28). The meaning of a free combination is purely literal. That is, the whole is the sum of its parts. Two examples of free combinations are *drink tea* and *under the table*. Collocations (also called restricted collocations) lie between these two extremes of the continuum. These semi-fixed combinations consist of elements possibly substitutable, however, with certain restrictions. Although a collocation may contain both literal and figurative elements, its complete meaning is transparent (Nesselhauf, 2005). In the collocation *throw a party* for example, *throw* is used in a figurative sense and *party* in a literal sense. However, the meaning of the whole expression is straightforward. The verb *throw* is possibly replaced by verbs like *give* and *have* but not many others. In analyzing a collocation, the element being focused on is called a 'node' and its possible combinational partners called 'collocates'. The number of potential collocates a word can take is defined as 'collocational range' or 'collocability' (Cowie, 1981; Handl, 2008).

Figure 2.1. The continuum of idiomaticity (Adapted from Howarth, 1998, p. 28)

Fourth, collocations are possibly stored and processed like single lexical items. Many collocation researchers believe that collocations, although analyzable into parts, are acquired and stored as prefabricated units in the mental lexicon (e.g., Bolinger, 1976; Palmer, 1933; Pawley & Syder, 1983; Sinclair, 1991; Wolter & Gyllstad, 2011; Wray, 2002). For example, Wolter (2001) argues, in his Depth of Word Knowledge Model, that collocational connections go beyond the knowledge of individual words and that such connections are built in an intermediate stage in the development of metal lexicon. The assumption that collocations are whole units has been supported by accumulating evidence from psycholinguistics research which indicates that formulaic language such as collocation is processed more efficiently than generated language by both native and non-native speakers (see e.g., Conklin & Schmitt, 2008; Schmitt & Underwood, 2004; Underwood, Schmitt, & Galpin, 2004; Wray, 2002). Additionally, word association (WA) research has suggested that collocational associations play an essential role in the organization of first language (L1) and L2 mental lexicons (e.g., Fitzpatrick, 2006; Fitzpatrick, 2007; Namei, 2004; Wolter, 2001; Zareva, 2007).

2.2 Collocational Knowledge and General L2 Proficiency

> [Words] are interconnected, not isolates … meaning is derived from context, and … collocation is key (Moon, 2008, p. 243).

Moon's words aptly describe collocation's scaffolding role in language. There is almost consensus among language researchers, particularly vocabulary researchers, that collocation is pervasively used in all languages (e.g., Ellis, 2008; Nesselhauf, 2005; Schmitt, 2010; Wray,

2002). For this reason, the argument that being able to process collocations effectively and efficiently in language reception and production is a defining aspect of language proficiency (Luoma, 2004; Nesselhauf, 2005; Towell, Hawkins, & Bazergui, 1996; Wray, 2002) seems highly plausible.

Although collocations are fundamental building blocks of a language, L2 learners are found to lack adequate knowledge of them. A large body of research has focused on collocation use in L2 writing. It is found that L2 learners generally use fewer collocations than native speakers when writing essays (e.g., Granger, 1998; Herbst, 1996; Howarth, 1996; Laufer & Waldman, 2011; Sugiura, 2002; Zhang, 1993) and that the collocations they produce is prone to error (e.g., Herbst, 1996; Howarth, 1996; Laufer & Waldman, 2011; Nesselhauf, 2003, 2005; Siyanova & Schmitt, 2007, 2008). In addition, it seems that L2 learners tend to overuse collocations they know well at places where more precise language is needed (Durrant & Schmitt, 2009; Shih, 2000; Sugiura, 2002).

Previous second language acquisition (SLA) research has suggested a positive relationship between the amount of L2 collocational knowledge and the level of general L2 proficiency (Al-Zahrani, 1998; Bonk, 2000; Gitsaki, 1999; Hsu, 2007; Keshavarz & Salimi, 2007). It is found that high-level L2 learners usually perform better than low-level L2 learners in well-designed collocational knowledge tests (Al-Zahrani, 1998; Gitsaki, 1999; Hsu, 2007; Voss, 2012; Zughoul & Abdul-Fattah, 2003). Notably, Hsu (2007) observed a clear developmental pattern of collocation production in sixty-two Chinese ESL students' writing. The quantity, accuracy, and diversity of collocation production increased with the writers' L2 proficiency.

However, it is hard to combine the results of the aforementioned studies because of their inconsistent designs (Table 2.1). First, these studies focused on various collocation patterns

which may differ widely in difficulty. For example, grammatical collocations may be easier to translate than lexical collocations (Gitsaki, 1999) while among lexical collocations, the adverb-adjective pattern may be the easiest to learn (Martyńska, 2004). Second, these studies assessed collocational knowledge using different methods. Some used receptive measures (e.g., multiple-choice, matching, error recognition, and acceptability judgment) while others chose to elicit and evaluate constructed responses (e.g., translation and free writing). Testing methods, according to some language testing researchers (e.g., Douglas, 1998; Shohamy, 1984), may affect both the intended test constructs as well as the observed test performance. Zughoul and Abdul-Fattah's (2003) study, for example, suggests that L2 learners tend to perform better in receptive than productive collocation tasks, indicating that the two task types assess different aspects of collocational knowledge. Leśniewska and Witalisz (2007) found that translation tasks caused more cross-lingual errors (i.e., errors caused by translating L1 to L2) in learner language than other collocation elicitation tasks. The testing methods in this study seemed to encourage language use in certain way. Third, the participants of these studies were from various L1 backgrounds. The linguistic distance between L1 and L2, according to Nesselhauf (2005), is a strong predictor for the difficulty in learning L2 collocation. Specifically, L2 expressions that are congruent with L1 translations tend to be easier to learn than those that are not.

Unfortunately, none of the above studies adopted oral proficiency scores as a criterion measure. As the construct of speaking may diverge significantly from the constructs of listening, reading, and writing (see e.g., Huff et al., 2008; L. Wang, Eignor, & Enright, 2008), it is unclear whether and to what extent L2 collocational knowledge relates to L2 oral proficiency.

Table 2.1 A Summary of the Studies on the Relationship between L2 Collocational Knowledge and General L2 Proficiency

Authors	Participants	Syntactic Pattern	Test method	Criterion of L2 proficiency
Al-Zahrani (1998)	Arabic learners of English	V-N	Cued cloze test	Level of education, writing test scores, and paper-based TOEFL scores
Bonk (2000)	Asian learners of English	V-N and V-P	Cloze test and acceptability judgment	Paper-based TOEFL scores and length of residence
Gitsaki (1999)	Greek learners of English	26 grammatical and 11 lexical	Free writing, cloze test, and translation test	Level of education
Hsu (2007)	Taiwanese learners of English	V-N, ADJ-N, N-V, N-of-N, ADV-ADJ, V-ADV, and N-N	Free writing	TOEFL writing scores
Keshavarz and Salimi (2007)	Iranian learners of English	V-N, ADJ-P, N-P, and V-P	Multiple-choice	Cloze test scores
Martyńska (2004)	Polish learners of English	V-N, ADJ-N, N-N, N-V, ADV-ADJ, N-of-N, and V-ADV	Matching, translation, cued gap-filling, multiple-choice, and error recognition	Length of English learning experience
Nizonkiza (2011)	Burundian learners of English	sampled based on frequency	Word association	Paper-based TOEFL scores and Level of education
Zughoul and Abdul-Fattah (2003)	Arabic learners of English	16 *kasara*-collocations (i.e., Arabic verb 'broke')	Multiple-choice and free translation tasks	Level of education

2.3 Collocation and Speaking

A theory for collocational language ability as an important subcomponent of the construct of L2 oral proficiency would provide strong support for adding a collocation evaluation component into SpeechRater (i.e., backing for assumption 1, see section 1.2 above). In addition, basing collocation scoring criteria on the assumed role that collocation plays in speech processing "has the potential to move from scoring as a kind of behavioral checklist … to a richer level in which the score provides insight into the mental processes underlying the performance" (Clauser et al., 2002, p. 428).

This section reviews four widely recognized speech-processing models in the literature so as to construct a preconceived theory about the relationship between collocation and speaking. These models include Bygate's (1987) model of speaking as a process, Levelt's (1989, 1999a, 1999b) modular model of speech production, de Bot's (1992) bilingual production model, and Kormos' (2006) bilingual speech production model. It is notable that these models are not independent from each other. Rather, strong connections can be noticed among them, reflecting the evolution of researchers' thinking on the construct of oral proficiency. The section will conclude with a logical analysis of the function of collocation in L2 speech production.

2.3.1 Bygate's Model of Speaking as a Process

Bygate's (1987) model of speaking as a process was developed for language teachers' better understanding of the nature of L2 speaking and for their pedagogical planning of oral classroom activities. In this model (Figure 2.2), speaking is viewed as a speaker-internal

process which involves the use of knowledge and skill in three consecutive processing stages: planning, selection, and production.

In line with Skill Acquisition Theory (DeKeyser, 2007), Bygate makes a distinction between knowledge and skill in spoken language ability. He refers to *knowledge* as knowing about a language (including vocabulary, a set of grammar and pronunciation rules, and the routines of using them) and *skill* as the ability to implement the knowledge in the speaking action. Although Bygate emphasizes the importance of skill over knowledge, he suggests an interdependent relation between the two. That is, skill is contingent upon the possession of appropriate knowledge resource and, in turn, the exercise or practice of skill helps expand the knowledge store.

Bygate's model starts with the planning stage in which a speaker decides on the topic, message, and how the information will be delivered. First, the speaker draws on the knowledge of routines to make plans of speech so that it conforms to the norms of language use (called 'message planning'). Routines are "conventional ways of presenting information" (Bygate, 1987, p. 23) and are further divided into two types: *information* and *interactional*. Information routines are frequently reoccurring information-based structures such as stories, instructions, or descriptions. Interactional routines are typical situation-based turn structures such as telephone conversations, interview situations, or casual encounters. In addition, the speaker needs to rely on his or her knowledge of the discourse structures to manage the interaction with interlocutors (called 'management skills'). Specifically, the speaker manages the agenda of the conversation (i.e., deciding on how topics are developed, when to maintain or change a topic, or when to end a conversation) and handle turn-taking (e.g., signaling the desire to speak, recognizing the moment to get a turn, or letting someone else to have a turn).

At the second selection stage, the speaker chooses expressions to communicate intended ideas based on his knowledge of lexis, phrases, and grammar resources. The skills associated with this stage deal with meaning negotiation with the interlocutors. On the one hand, the speaker needs to determine the level of explicitness of the information according to a listener's prior knowledge of the topic (called 'explicitness skills'). On the other hand, the speaker needs to follow procedures to control the level of specificity of his language use (called 'procedural skills'). That is, he has to decide when to use general terms and when to modify the precision of language use through metaphors, paraphrases, or emphases. The skills in the planning and selection stages all belong to the 'interactional skills' which reflect the reciprocal nature of the speaking activity. That is, the speaker has to constantly adapt speech based on the listener's reaction.

In the third production stage, the speaker orally produces the ideas planned and the language selected. In this process, he applies 'production skills' and 'accuracy skills' to maintain the fluency and accuracy of speech. The production skills include *facilitation* and *compensation*. Facilitation refers to the effective use of production devices to speak economically. It helps the speaker, especially a non-native speaker, to lighten his cognitive load while speaking under time pressure. Facilitation can be achieved in four main ways: simplifying structure, ellipsis, using formulaic expressions, and using fillers and hesitation devices. Compensation, on the other hand, is used to correct or improve some aspects of the messages previously said. Some typical compensation skills include self-correction, false starts, repetition, and rephrasing. According to Bygate, such correction behavior is forgivable and even necessary in consideration of the limited planning time allowed for an impromptu speech. Finally, the accuracy skills guarantee that the speaker speaks in an accurate manner

based on his knowledge of the grammatical and pronunciation rules. According to Bygate, the production and accuracy skills alone are not adequate for fluent and accurate oral production; it is "the ability to handle all the sub-skills from the top of the diagram to the bottom" that accounts for the speaking ability (p. 49).

In addition to delineating the process of speech production, Bygate identifies two major types of compensatory strategies that L2 speakers commonly take to make up for their language deficiency: *achievement* and *reduction* (see also Fulcher, 2003 for a similar discussion). By using the achievement strategies, L2 speakers are able to fill in the lexical gaps with substitutes from their existing language resources. They may *guess* an unknown word or expression by borrowing or literally translating an equivalent from their native language or creatively coin a word based on their knowledge of the target language. Alternatively, they may *paraphrase* the gap using a rough synonym (called *circumlocution*, i.e., to use an unnecessarily large number of words to express an idea). The deviant expressions in L2 speech, according to Bygate, mainly result from the use of the achievement strategies. The reduction strategies, on the other hand, refer to the strategic abandon of the original communicative objective because of the lack of language resources. In order to stay out of trouble, L2 speakers may avoid words with difficult pronunciation or expressions with complex grammatical structures. Alternatively, they may completely give up a planned message and confine themselves to manageable topics (e.g., false starts). Swain, Huang, Barkaoui, Brooks, and Lapkin's (2009) investigation into TOEFL iBT test takers' self-reported strategic behaviors in L2 speaking provided empirical evidence for Bygate's theory of compensatory strategies.

In Bygate's speaking model, the ability to use formulaic expressions is part of the facilitation skills at the production stage. He emphasizes their importance to cognitive economy and oral fluency:

> Our interest in these expressions is that they can contribute to oral fluency. Speakers do not have to monitor their choice of words one after another. They do not have to construct each new utterance fresh, using the rules of the grammar and their knowledge of vocabulary in order to vary their expression for each afresh occasion. Instead they proceed by using chunks which they have learnt as wholes. This is particularly important in routine situations (Bygate, 1987, p. 17).

Bygate's definition of formulaic expressions seems to embrace a large proportion of collocations. He describes the term as "conventional 'colloquial' or idiomatic expressions or phrases" or "set expressions" which include not only idioms but also "phrases which have more normal meanings" and "tend to go together" (Bygate, 1987, p. 17). Obviously, collocations, particularly those which are highly restricted to element substitution, fall in this category.

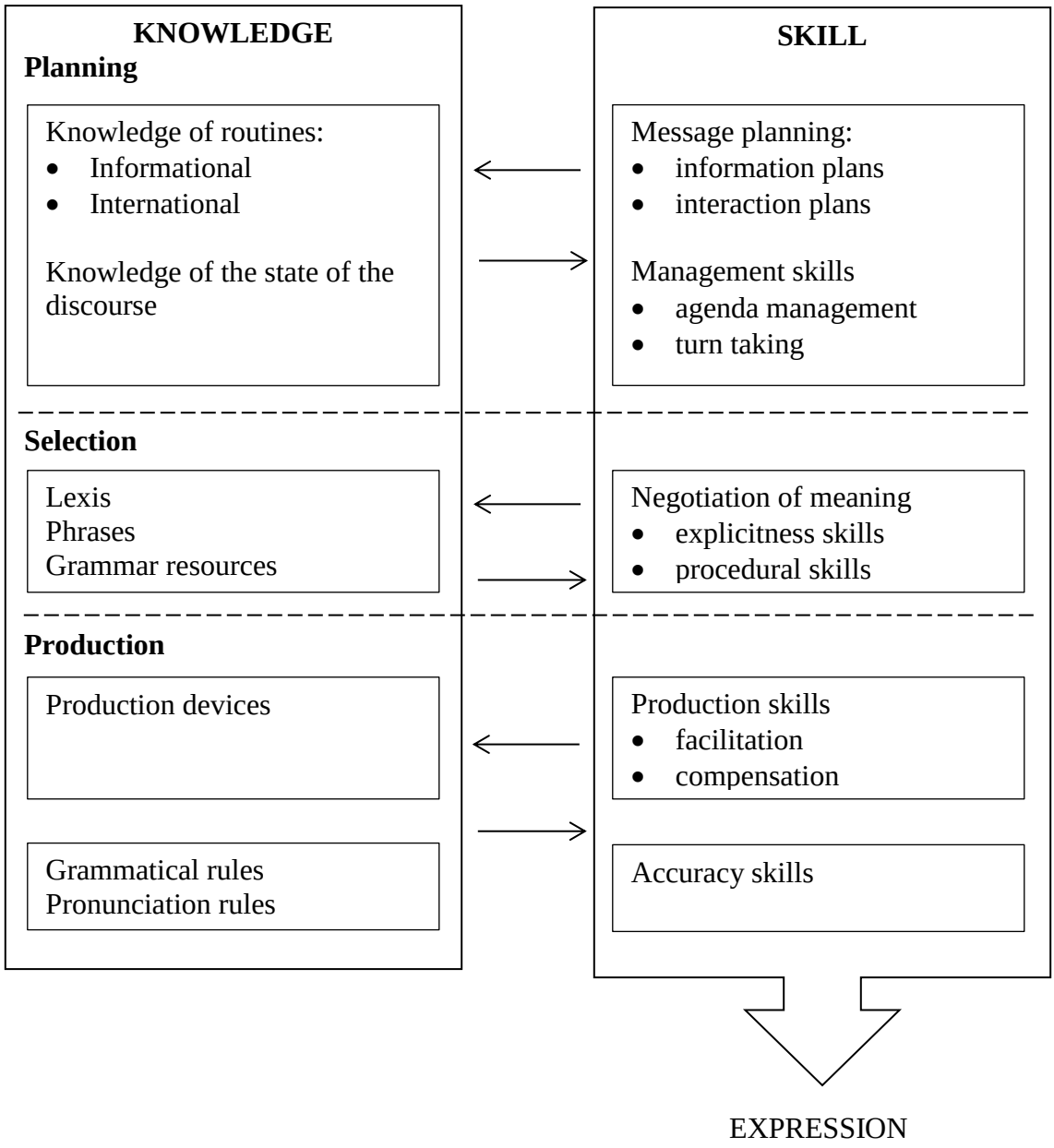


Figure 2.2. A model of speaking as a process (Bygate, 1987, p. 50)

2.3.2 Levelt's Modular Model of Speech Production

Levelt's (1989, 1999a, 1999b) modular model (Figure 2.3) might be the empirically best supported and the most influential model for monolingual speech production (Derwing, Munro, Thomson, & Rossiter, 2009; Kormos, 2006). His model seems to be deeply

influenced by the research on the development and maturation of an L1 child's speaking ability and that on speech error and chronometry of naming (i.e., the reaction time of naming objects). Levelt argues that human speech production is realized via two core systems which emerge from evolution: a *rhetorical/semantic/syntactic* or meaning-making system, which is responsible for mapping the message a speaker intends to express onto language representation, and a *phonological/phonetic* or articulatory system, which is in charge of the oral production of the language representation (diagramed as big rounded rectangles). A fluent language speaker is supposedly capable of coordinating these two underlying systems effortlessly but there might still be rifts between them (Levelt, 1999b). Each core system is further composed of a series of highly autonomous or modular information processing components, including *conceptual preparation*, *grammatical encoding*, *morpho-phonological encoding*, *phonetic encoding*, and *articulation* (diagramed as small rectangles), through which speech is formulated from intention to utterance.

Levelt assumes that the functioning of the two core systems relies on three knowledge stores (diagramed as ellipses), which are akin to the knowledge resources in Bygate's model. The first knowledge store is called *knowledge of external and internal world*. It comprises the *model of addressee* or Theory of Mind (ToM), which is the speaker's awareness of the social environment (e.g., who the interlocutors are and how much they know), a *discourse model* containing the speaker's diligent bookkeeping of entities (e.g., events, persons, and so forth) to which reference can be made as the discourse situation continuously changes, and *encyclopedic knowledge*, i.e., information about the world. By referring to this knowledge, a speaker determines what to say given a speaking context. The second knowledge store, namely *mental lexicon*, is accessed by the speaker to encode concepts into language. As its

name suggests, mental lexicon is a store of semantic, syntactic, morphological, and phonological information about the words in a language. Levelt assumes that the same mental lexicon is used for both constructing and perceiving speech[1]. The third store named *syllabary* is a repertoire of the articulatory-motor gestures of many high-frequency syllables in a language. It is drawn upon by the speaker when preparing his or her *internal speech*, a phonetic plan made for the actual articulation.

In Levelt's multi-stage model, speaking starts from *conceptual preparation* in the *rhetorical/semantic/syntactic* system. In this very beginning phase, a speaker generates a *message* that contains the conceptual information to realize his certain communicative intention. The conceptual generation involves two sub-processes: *macroplanning* and *microplanning*. Through macroplanning, the speaker decides on discourse focus: what information to convey and in what order. Microplanning, on the other hand, allows the speaker to further elaborate the information to make it more expressible in language. The elaboration includes deciding on perspective taking, estimating the degree of accessibility of the referents (to objects, persons, situations for example) to the addressee, marking certain information as prominent to draw the addressee's attention, using various devices to guide the addressee's attention, and translating spatial or imagery information (e.g., the mental image of a scene) into language-specific propositional format or "sets of relations holding between concepts" (Levelt, 1989, p. 72). The product of the macroplanning and microplanning processes is a *preverbal message* which is a "conceptual structure consisting of lexical concepts" (Levelt, 1999b, p. 87). In other words, these concepts are organized in such a way that they are ready to be converted into language.

[1]Bachman and Palmer (1996, 2010) seem to adopt some ideas from Levelt's conception of knowledge stores in conceptualizing a universal language use model that is not limited to speaking. They call *encyclopedic knowledge* 'topical knowledge' and the *mental lexicon* 'language knowledge' instead.

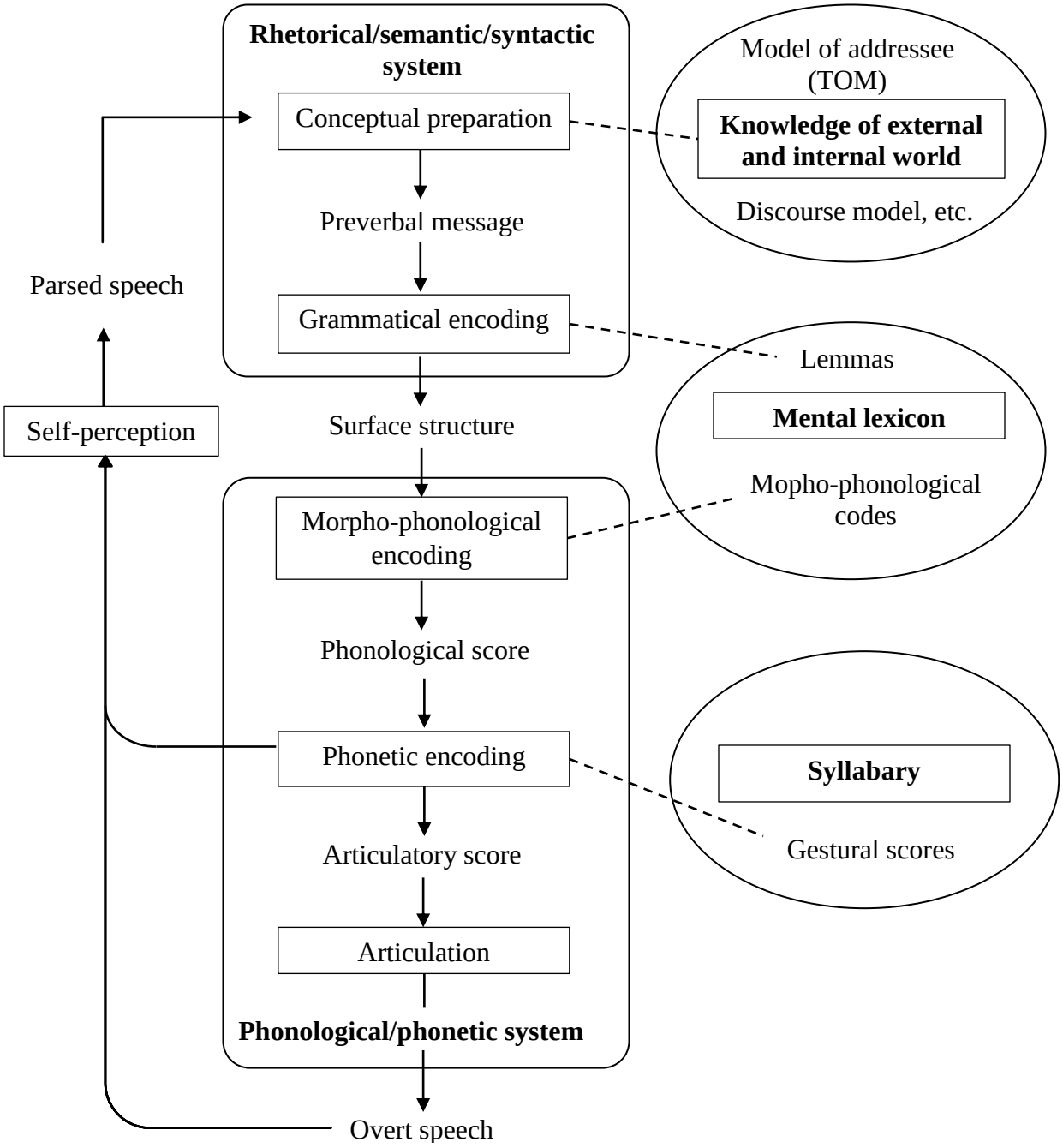


Figure 2.3. A modular model of speech production (Levelt, 1999b, p. 87)

The preverbal message is then passed to the next processing component, called *grammatical encoding*. At this point, the speaker constructs a grammatical *surface structure* onto which the lexical concepts and their relations in the preverbal message are mapped. Grammatical encoding again contains two sub-processes: *lemma selection* and *syntactic composition*. Lemma selection refers to the speaker's act of retrieving 'lemmas' (also called syntactic words, such as nouns or verbs) from the mental lexicon, whose meaning and sense, including pragmatic, stylistic, and affective features, correspond to the lexical/semantic concepts in the preverbal message. Lemmas[2] are an immediate form of words which contain the syntactic and morpho-phonological information that is useful for subsequent syntactic composition and morpho-phonological encoding phases. Levelt considers this selection procedure as a "probabilistic affair" (Levelt, 1999b, p. 96). That is, the target lemma is selected from a group of lemmas that correspond to the intended concept as well as many semantically related concepts. For example, when having a lexical concept *sheep* in mind, the speaker needs to choose among three lemmas *sheep*, *goat*, and *llama* that are activated at the same time in the mental lexicon. According to Levelt, it is the competition among these nearly synonymous lemmas that slows down the selection process. When the lemmas are chosen, the syntactic composition is initiated. This is a *unification* process in which the speaker connects lemmas into syntax or surface structure, which is "an ordered string of lemmas grouped in phrases and subphrases of various kinds" (Levelt, 1989, p. 11). The generation of syntax is incremental. That is, as soon as certain lexical concepts or even fragments of them are ready, their corresponding lemmas are selected and the unification

[2] De Bot (1992) made a distinction between lemma and lexeme: A lemmas contains a lexical entry's "meaning and syntax" whereas a lexeme keeps its "morphological and phonological properties" (p. 4). Hence, a lemma is activated before the corresponding lexeme in the mental lexicon for selecting the right words.

begins. For this reason, the resulting surface structure is to some extent determined by the order in which lexical concepts become available (Levelt, 1999b).

The surface structure serves as the input for the second *phonological/phonetic* system in which articulation takes form. Once the lemmas are chosen and positioned in the syntax, their morphological and phonological information, including morphological makeup, metrical shape, and segmental makeup, are activated and become available in the mental lexicon. The speaker retrieves this memory information in the course of *morpho-phonological encoding* to generate a *phonological score*, which is composed of phonological syllables as basic units and is organized in a hierarchy of phonological words and phrases. In this process, the discrete morpheme segments of the lemmas (e.g., /s/, /i/, /l/, /ɛ/, /k/, /t/, /ʌ/, and /s/ for the phrase 'select us') are spelled out and then incrementally grouped into syllables based on the prosodic context (e.g., /si/, lɛk/, and /tʌs/ for the phrase 'select us'). In addition, the stressed syllables are assigned and pitch movements or intonation patterns are determined (for more detail, see Levelt, 1999b). The phonological score is also called *internal speech* and is the intermediate internal representations that the speaker monitors before the articulation.

In the subsequent *phonetic encoding* phase, the syllables in the phonological score are realized through articulatory gestures, the motor routines that a person develops in three articulatory tiers (i.e., glottal, nasal, and oral) in the process of language learning. The speaker rarely composes entirely new syllables but rather have easy access to the *ready-for-use* gestural scores in the *syllabary* for frequently used syllables of the language. The output of *phonetic encoding* is called articulatory score which is an abstract entity concerning various parameters in articulation such as tongue position, duration, pitch movement, key,

register, and so on. In the final *articulation* stage, the articulatory score is executed by the laryngeal and supralaryngeal apparatus to produce the ultimate product: overt speech.

Levelt also hypothesizes an additional self-monitoring component in the model. The component is called a *self-perception* system through which the speaker continuously monitors his *internal speech* or the overt speech in order to maintain the accuracy and appropriacy in oral production. Levelt calls this feedback mechanism the 'perceptual loops' and named three of them.

The first one is called a conceptual loop through which a speaker can directly monitor the preverbal message that contains conceptual information in the *semantic/syntactic* system either before or after it is sent for syntactic composition. For example, the speaker may reject a message after its formulation has started. This is observed as a false start (e.g., 'Tell me, uh what—d'you need a hot sauce?') in discourse that the speaker formulates the initial portion of the speech but abandons it before finishing (Du Bois, Schuetze-Coburn, Cumming, & Paolino, 1993). One of the reasons that a speaker modifies a preverbal message is that he finds out that the original message is no longer necessary, appropriate, or acceptable in the given communicative situation (Kormos, 2006).

The second loop allows the speaker to monitor his pre-articulatory, internal speech in the *phonological/phonetic* system. Specifically, the speaker checks on possible errors in the syllabified phonological words before actually articulating them. However, this phonetic/phonological self-monitoring is not used exclusively for detecting word form errors. The pre-articulatory loop may start at the phonological level of processing but can "rush forth into the syntactic/semantic domain" when needed (Levelt, 1999b, p. 114).

The third monitoring mechanism occurs after articulation. The speaker can hear errors in his actual speech via the speech perception system that he relies on to parse or decode others' speech. It is therefore called an external loop and is considered the most likely route of self-monitoring (Levelt, 1999b). When the speaker notices errors or inapproriacy in any of the above three loops, he can interrupt the utterance and make corrections immediately.

A claim that Levelt has constantly made in his works is that the speech encoding process is incremental rather than strictly serial as in an assembly line. That is, any autonomous processing component in the model does not idle or wait for the previous one to fully complete its job. Instead, each module can be triggered into action by any fragment of input provided by the previous module. In addition, a fragment can be processed independently "without much look ahead" (Levelt, 1989, p. 24). In other words, a processor can successfully encode a fragment without referring to other fragments that it expects to receive. This properly explains why a speaker's utterance of the first few words does not depend on how the sentence will be finished. The ability to deal with incomplete information input at each processing stage grants the entire hypothesized speech encoding model the maximum multitasking power. With such abilities, all the processing components can run in parallel, "overlapping their processing as the tiles of a room" (Levelt, 1999b, p. 88). When a speaker is articulating a phrase, he is already organizing the content for the next phrase. This explains why the extremely demanding task of speaking can be performed so elegantly.

2.3.3 De Bot's Bilingual Production Model

De Bot (1992) adapted Levelt's (1989) unilingual speaking model to account for bilingual or multilingual speaking behaviors. The adapted model is able to explain four

bilingual/multilingual language phenomena: independent use of L1 or L2 as well as the code-switching between the two, cross-linguistic influences, unequal proficiency in L1 and L2, and the interaction between different language systems. The structure of de Bot's model is displayed in Figure 2.4. A detailed description of it is unnecessary and possibly confusing because the adaption is made based on Levelt's earliest model.

A critical issue that de Bot had to resolve is whether bilingual speech production occurs in two entirely separate systems. Slightly disagreeing with Levelt, he argues that language-specific encoding starts in the *microplanning* rather than the *macroplanning* process in the *conceptualizer* (the predecessor of the *conceptual preparation* component in Levelt's latest model). In other words, the concepts a speaker formulates in the first place are not different between languages; it is when the speaker organizes the propositional content (e.g., argument structure, referents, mood, etc.) that the speech formulation starts to split into two language systems (see Figure 2.4). This separation, according to him, will remain in the following *grammatical encoding* and *phonological encoding* components[3] to generate language-specific phonetic plans (internal speech), which will eventually be sent to the articulator to produce overt speech.

[3] In the early version of Levelt's model, the morpho-phonological encoding and the phonetic encoding components are not distinguished. Together, they are called a phonological encoding component.

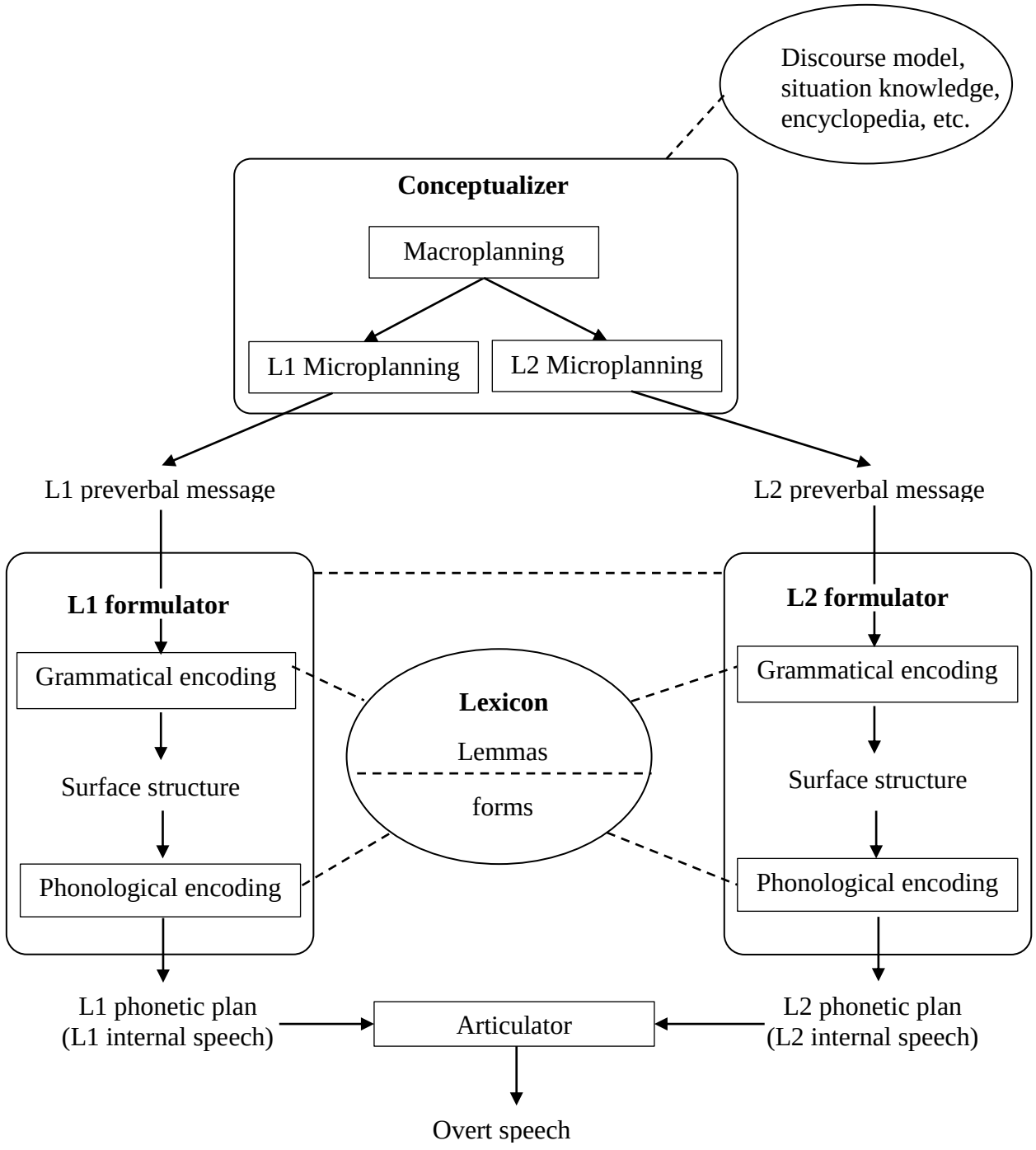


Figure 2.4. A bilingual production model based on de Bot's (1992) description

De Bot posits the existence of a single lexicon to account for bilingual or multilingual speakers' coding-switching behaviors. According to him, the lexicon keeps all the semantic, syntactic, morphological, and phonological information of the lexical items in all the languages that the speaker masters. This information is retrieved at the formation of the surface structure and the phonetic plan (i.e., internal speech). As a follower of Paradis' (1981, 2004) Subset Hypothesis, de Bot believes that the lexical items are connected in networks and that those belonging to the same language are more strongly connected and thus can be activated together. In a balanced bilingual speaker's lexicon, however, the lexical connections in one language system are as strong as those across language systems. This explains why such speakers can code-switch smoothly between two or more languages. This idea of a shared lexicon among languages was later adopted by other bilingual speaking models (e.g., Hartsuiker, Pickering, & Veltkamp, 2004; Kormos, 2006). A further discussion of the organization of the bilingual lexicon can be found in de Bot (2004).

2.3.4 Kormos' Model of Bilingual Speech Production

Kormos' (2006) model of bilingual speech production is an update on Levelt and de Bot's models reviewed earlier. Compared to the two previous models, it is more suited to explain the difficulties of speaking L2. In addition, Kormos emphasizes the indispensable role of formulaic language to speaking more strongly.

Like Levelt and de Bot, Kormos posits that speech encoding is modular and incremental and that it involves retrieving information from some knowledge stores. Her speaking model is made up of a *conceptualizer*, a *formulator*, an *articulator*, a *long-term memory* bank, a *speech-comprehension* system, an *audition* component, and three monitoring

loops that deal with preverbal (conceptual) message, internal speech, and overt speech, respectively (Figure 2.5). She argues that monitoring L2 speech production requires more attentional resources than monitoring L1 speech production because the former is less automatized and more error-prone than the latter (see more discussion below).

In Kormos' model, a single formulator is assumingly responsible for encoding multiple languages. This reflects a simplification on de Bot's model. According to her, the simplification is motivated by the principle of ecology in human cognition as well as the recent neuroimaging research that reveals the similar nature between L1 and L2 speech processing.

Based on the mainstream theories of memory, Kormos proposes the existence of a single memory bank that holds all the conceptual and linguistic resources. This memory bank, namely *long-term memory* (diagramed as a large oval), contains four distinctive components: episodic memory, lexicon, syllabary, and a store for declarative knowledge of L2 rules. The episodic memory stores personally experienced events along with the emotions associated with them. The lexicon, which is drawn upon in both speech encoding and comprehension (connected in dotted lines), keeps a structured record of linguistic and non-linguistic concepts in three levels: the conceptual knowledge that is accumulated over the years, the lemmas that contain syntactic information, and the lexemes that contain morpho-phonological information. According to Kormos, the episodic memory and the lexicon are closely related. That is, a person learns concepts and language through personal experience and the acquired language knowledge can trigger the recollection of the specific experience of learning. The syllabary, as also seen in Levelt's model, stores automatized articulatory-motor gestures to produce syllables in L1 and L2. The store of L2 declarative rules is a

"declarative memory of syntactic and phonological rules in L2" (Kormos, 2006, p. 167). It is solely used for L2 speech encoding. However, using these rules supposedly consumes a great deal of attentional resources.

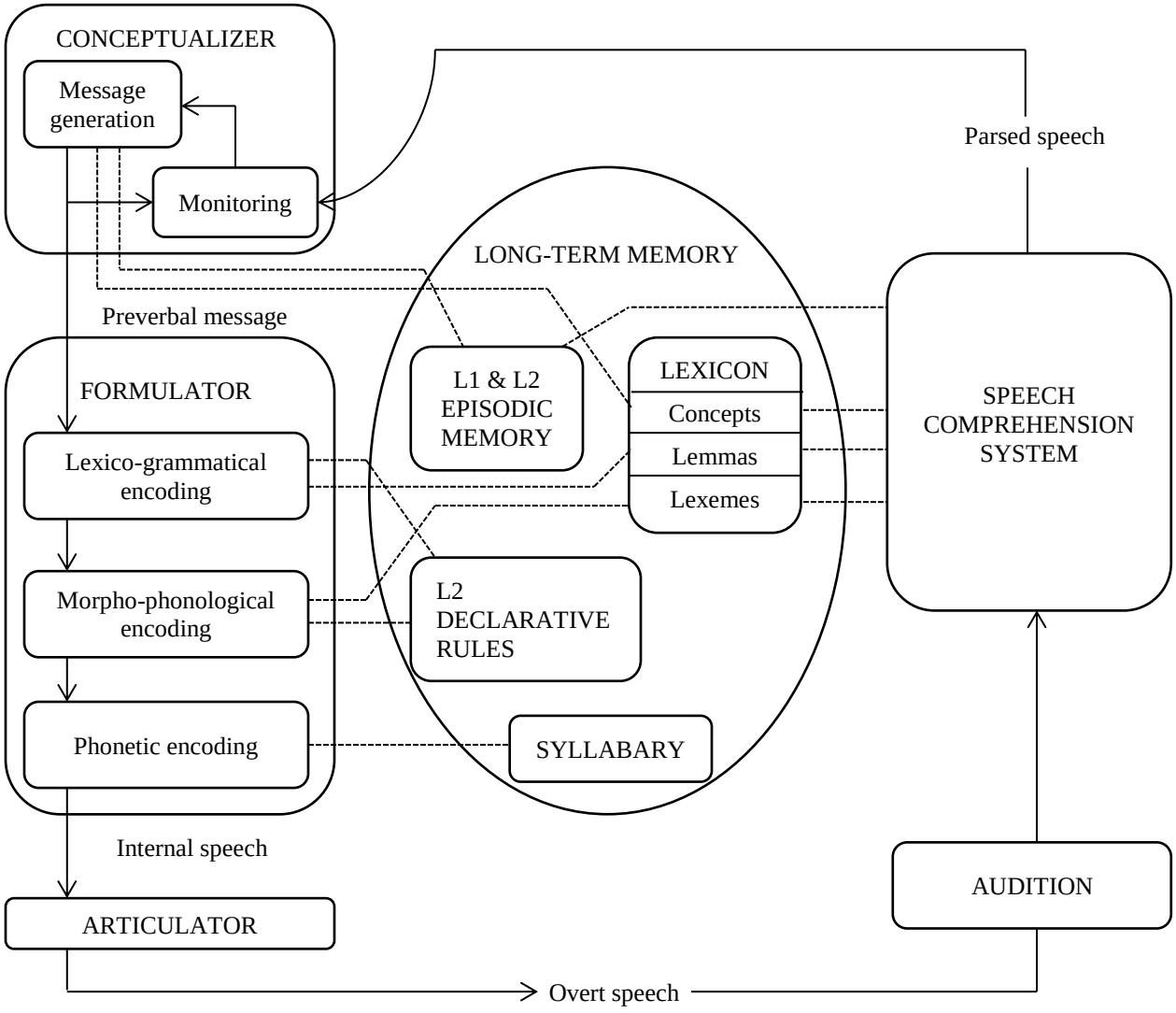


Figure 2.5. A model of bilingual speech production (Kormos, 2006, p. 168)

Kormos relates the development of L2 oral proficiency to a number of changes in the constructs of her model. In her theory, highly proficient L2 speakers resort less to the store of

L2 declarative rules but more to the lexicon. This is because advanced learners' declarative knowledge has been largely transformed into procedural knowledge in the lexicon, which can be retrieved at minimum cost of attentional resources; a means of this transformation, according to her, is the memorization of frequently used formulaic expressions.

Kormos echos Bolinger's (1976) opinion that formulaic expressions are key elements in the spoken language. In her words, "The majority of our utterances are combinations of memorized phrases, clauses, and sentences, which together are called formulaic language" (Kormos, 2006, p. 170). She posits that the proficient use of the formulas starts from conceptualizer's retrieval of the *conceptual chunks* from the lexicon that are associated with particular communicative functions (p. 45). As these conceptual chunks are closely mapped onto the *lexical chunks* in an advanced speaker's lexicon, the formulator will be able to retrieve and activate the lexical chunks easily. Chunking, according to Kormos, contributes enormously to speech automaticity (p. 42). And the chunks in Kormos' terminology not only refer to fixed expressions such as idioms or conventionalized expressions but to semi-fixed collocations as well. She believes that the knowledge of chunks is proceduralized through the increasing exposure to the target language and practice of using the language (see also Durrant & Schmitt 2010; Wolter & Gyllstad, 2011).

As regards the L1 influence on L2 speech production, Kormos gives two plausible explanations. One is that L2 speakers may choose an L1 item or rule by mistake although they have acquired the L2 equivalent. In Kormos' model, the L1 and L2 concepts, lemmas and lexemes are stored together in the lexicon; for this reason, they compete for selection in the speech encoding process. According to Kormos, a mistake in selection is likely to occur

when the L1 item or rule shows a higher level of activation than the L2 counterpart. This is believed to happen more often among unbalanced bilinguals who use L1 more frequently.

The other explanation concerns L2 compensatory strategies (see also section 2.3.1 above). That is, speakers who are less proficient may assume that L2 works in the same way as L1, hence drawing on L1 knowledge to fill L2 gaps. Different from de Bot (1992), Kormos holds that L1 and L2 concepts do not completely overlap in the lexicon; some concepts, depending on how and where they were acquired, may be encoded in a specific language. For example, the concept of 'family reunion' may evoke totally different mental images and memories between two cultures (Xu, 2010). Based on this assumption, Kormos defines two types of L1 transfer. She refers to *semantic transfer* as an association between an L2 lemma with an L1 concept (when L1 and L2 concepts do not seamlessly overlap) and *syntactic transfer* as an association between an L2 lexeme with the syntactic information of its L1 counterpart (when L1 and L2 concepts overlap).

2.3.5 Logical Analysis of the Relationship between Collocation and L2 Oral Proficiency

This dissertation study sought a plausible construct theory to explain and support the hypothesized tight relationship between collocation and L2 oral proficiency. Although the contemporary speaking theories tap into formulaic language, they stop short of elaborating on the contribution of collocation to speech formulation. In this section, I applied "logical attack" (APA, 1954, p. 54) to establish (logical reasoning) validity of the above hypothesis, making logical connections among the construct of collocation, speech-processing theories, and L2 speech evaluation criteria.

A review of the speech-processing models has given us a clear picture of where an L2 learner may struggle with in spontaneous speaking. Based on these models, a speaker has to

coordinate multiple cognitive processes simultaneously under the time pressure, including situation and goal analysis, idea formulation, lemma selection, syntactic composition, articulation, and self-monitoring and correction. Thus, from a psycholinguistic perspective, the greatest challenge that every L1 or L2 speaker confronts is to allocate and make effective use of the limited attentional resources to achieve desired communicative goals (see 'Cognitive Load Theory', Chandler & Sweller, 1991).

To make matters worse, an L2 speaker assumingly has to devote extra attention to drawing on the learned L2 declarative rules, choosing and applying compensatory strategies, and monitoring the internal and overt speech (see Kormos, 2006, p. 173). In this regard, the crux of L2 speaking is to reduce the cognitive burden by proceduralizing or automatizing as many components of speech encoding as possible, just like native speakers do. Based on the trade-off hypothesis (Skehan, 1998, 2009), the attentional resources saved at one place can be spent on improving the other parts of speech performance. This view is in line with the Skill Acquisition Theory which claims that language learning is analogous to "the learning of a wide variety of skills" (DeKeyser, 2007, p. 97). If we follow this logic, then the link between collocation and L2 oral proficiency apparently starts from the basis that collocational ability is a kind of energy-saving skill that frees up mental resources during speaking.

L2 speaking assessment makes inferences about the L2 oral construct, which is usually deemed as a holistic concept, based on samples of a speaker's overt speech product. Some common analytic evaluation criteria for assessing the L2 oral construct include accuracy, fluency, speech intelligibility, content development, coherence, interaction, and so forth. Depending on the specific test purposes, different sets of criteria are chosen. However, there is almost a consensus that accuracy, fluency, and intelligibility are three fundamental

criteria of L2 oral proficiency (Fulcher, 2003; Luoma, 2004). I argue that strong collocational performance in speaking contributes to all these three aspects.

Using native-like collocations no doubt increases the accuracy of speech. Collocations support a speaker's ideas in the most economical way. Filling a collocational gap with an L1 translation, a rough synonym, or a paraphrase (Bygate, 1987; Fulcher, 2003; Schmitt, 1998) would certainly reduce the precision of vocabulary use. Wray and Fitzpatrick (2010) hold that using formulaic language reflects a speaker's intent "to select the most native-like expression from a larger set of ways in which a particular message might be grammatically expressed" (p. 37). Luoma (2004, p. 16) regards "well-chosen phrases" in L2 speech as evidence for the richness of the speaker's lexicon and suggests rewarding this aspect of language performance in the assessment.

Likewise, effective use of collocations supposedly improves L2 oral fluency and automaticity. Based on Kormos' (2006) theory, retrieving and encoding collocations as whole units saves a huge amount of mental resources and are thus more efficient. There is now convergent evidence for the notion that native speakers rely heavily on formulaic sequences to maintain fluency (Kormos, 2006; Schmitt, 2010; Wróbel, 2011). For example, their utterances of formulaic expressions often form uninterrupted intonation units, hence playing an important part in establishing speech rhythm (Lin, 2010; Millar, 2011; Van Lancker, Canter, & Terbeek, 1981). Even L1 speakers who suffer from aphasia still keep a repertoire of automatic utterances of conventional expressions (Van Lancker & Canter, 1981). In this regard, Schmitt's (1998) view seems to be right— "to some extent, the mind [of a native speaker] does seem to organize words according to their collocational links" (p. 28). Wood's studies indicate that formulaic language improves L2 oral fluency as well. He

found that using collocations extended the length of runs (the number of syllables) between pauses in learner speech (Wood, 2006) and that focused instruction on collocation increased learners' oral fluency to various degrees (Wood, 2007, 2009, 2010).

Finally, native-like collocation usage in L2 speech probably has a positive effect on native listeners' perception of the speech intelligibility and comprehensibility. Kormos (2006) posits that the speech comprehension system utilizes the same lexicon as the conceptualizer and the formulator. That is to say, native or highly proficient listeners may easily recognize the correctly used collocations in L2 speech and process them as whole units, thus saving the cognitive resources for speech analysis and decoding (see also Partington, 1998, p. 20; Millar, 2011; Voss, 2012, p. 76). There is growing evidence for the assertion that "formulaic language provides processing advantages over creatively generated (i.e. non-formulaic) language [based on L2 declarative rules]" (Schmitt, 2010, p. 136). However, the evidence was mainly obtained from reading comprehension research. According to Millar (2011, p. 142), "there has been little empirical evidence to show that the formulaicity of learner language *directly* [emphasis added] contributes to communicative competence".

2.4 The Construct of Spoken Collocational Competence

In light of the rationales given above, I quote Moon's (2008, p. 243) assertion once again, however, with a slight modification—"[C]ollocation is key" to assessing L2 oral proficiency. I propose formulating a new construct, namely spoken collocational competence (SCC) to account for L2 learners' collocational performance in naturally occurring speech. I also argue that SCC is the core component of the L2 oral construct.

First of all, I contend that SCC is related to but distinct from the construct of collocational knowledge (CK) that is assessed in various collocation tests (e.g., Voss, 2012). CK accounts for a language learner's performance in focused collocation processing such as judgment, recognition, translation, and written production when the time constraint is not so tight. That is, collocating typically involves performing a nearly exhaustive search in the lexicon using a great deal of attentional resources. In contrast, SCC accounts for the unfocused, procedural collocational performance in speaking. Such processing is performed at near-instantaneous speed and under none or little conscious control.

The distinction I make between SCC and CK is inspired by Wolter's theory about the organization of the bilingual lexicon (Wolter, 2001, 2006; Wolter & Gyllstad, 2011, 2013). Wolter postulates that an L2 speaker's bilingual lexicon consists of networks of L1 and L2 lexical items with well-known and frequently used items gathering at the center. The core items assumingly have more and stronger collocational connections than those in the periphery and are, generally speaking, more accessible. Based on Wolter's theory, SCC concerns the overall strength of the collocational connections in the central area of the network. CK, in contrast, concerns the strength of the collocational connections of the entire network, including the items in the periphery.

An extension I would like to add to Wolter's theory is that given the total number of lexical items stored in the lexicon fixed, there can be multiple L2 lexical networks, each corresponding to a domain of language use. The lexical items contained in each network are to some extent overlapping. However, the lexical items clustered in the center of each network (i.e., the most frequently used words and phrases) and the overall collocational strength among these items may vary enormously depending on an L2 learner's oral

proficiency in each domain of language use. This extension makes SCC a fluid and context-variant construct in accordance with the interactionalist perspective of construct definition (Bachman, 2007; Chapelle, 1998; see also Chapter 1, Section 1.2).

However, SCC would only remain an 'open concept' unless the construct is expressed as the function of observed variables (Pap, 1953; Whitehead & Russell, 1910). In other words, it is how SCC is measured in language assessment practice that assigns specific meanings to this hypothesized attribute. In this study, I conceive five subcomponents of SCC. They are semantic accuracy, grammatical accuracy, sophistication, transparency, and automaticity. I will give them more detailed explanation in Chapter 3, Section 3.2.

2.5 Automated Evaluation of Spontaneous L2 Speech

Because an important goal of this dissertation study was to inform the future development of a sophisticated automated scoring program for unrestricted automated speech evaluation (UASE), it is necessary to review the architecture and development agenda of SpeechRater, which, to my knowledge, is the only prototype of this kind of scoring program. Chapter 1 briefly mentioned the promise and main drawback of SpeechRater. This section will further elaborate on these issues and identify the major obstacles that hinder its development.

2.5.1 The Exciting Prospect of ASE

Learner language can prompt feedback as a marked-up text, a feedback message, or a score; it may be returned immediately or delayed. The feedback may be used for

> making high-stakes decisions, low-stakes decisions, or for informing students about the correctness of their language (Chapelle & Chung, 2010, p. 302).

L2 learners' speech production contains rich information about their oral language ability. Chapelle and Chung drew an ideal picture about how this information could be utilized to inform performance-based decision-making and autonomous language learning. Without technological assistance, it would be impossible to conduct such detailed analysis on learner language.

Expert human graders can only focus on a restricted range of speech features in oral language assessment. This is because the processing capacity of a human brain is constrained by the limited attentional resources (Chandler & Sweller, 1991). The Common European Framework, for example, indicates that "four or five categories [of scoring rubrics] begin to cause a cognitive load for raters and seven is a psychological upper limit" (Luoma, 2004, p. 80). Brown, Iwashita, and McNamara's (2005) study found that well-trained human graders only "attended to four major conceptual categories", with "the largest category of comments pertained to linguistic resources" (pp. 31-32).

Given the limitation of human grading, ASE gives us high hopes to provide instant and detailed feedback to language learners. A computer with enormous multitasking power is potentially capable of evaluating and generating feedback on every single construct-relevant element in L2 spoken data. In this respect, ASE will foreseeably have even wider construct coverage than human grading when all the technological constraints are removed. Such detailed diagnostic assessment feedback is greatly needed to understand learners' educational

needs (Chang, 2012) and to support language instruction and self-learning (Wolf, et al., 2014).

In addition, ASE can make oral assessment more efficient and accessible. As Xi (2010, p. 297) put it, "Computer capabilities, if used appropriately and responsibly, can expand the resources and improve the efficiency of language learning and assessment." This is particularly true in the context of large-scale speaking assessment in that rater training and test administration are logistically intensive and costly. A considerable delay in score reporting, according to Hauck, Wolf, and Mislevy (2013, p. 4), would have a negative impact on instruction and is one of the main drawbacks of the current generation of English language proficiency (ELP) assessments.

2.5.2 The Architecture of SpeechRater

SpeechRater is an automated scoring system designed by ETS to score low-stakes TOEFL iBT Speaking Practice Test (Xi, 2008). At present, it is also being used to monitor human rater performance, for example, by identifying extremely lenient and strict raters (Z. Wang, Zechner, & Sun, 2015).

According to Xi and her colleagues (Xi, et al., 2008), SpeechRater is made up of four modules: a speech recognizer, a number of feature extraction programs, a scoring model, and a user interface (Figure 2.6). First, the speech recognizer receives audio files of spoken data and converts them into words and utterances. Next, feature extraction programs extract construct-relevant scoring features from the words and utterances and send the features to a scoring model. The scoring model then evaluates these features to generate a composite test

score. Finally, a user interface reports the test score, the interpretations of the score, and test feedback to the test taker.

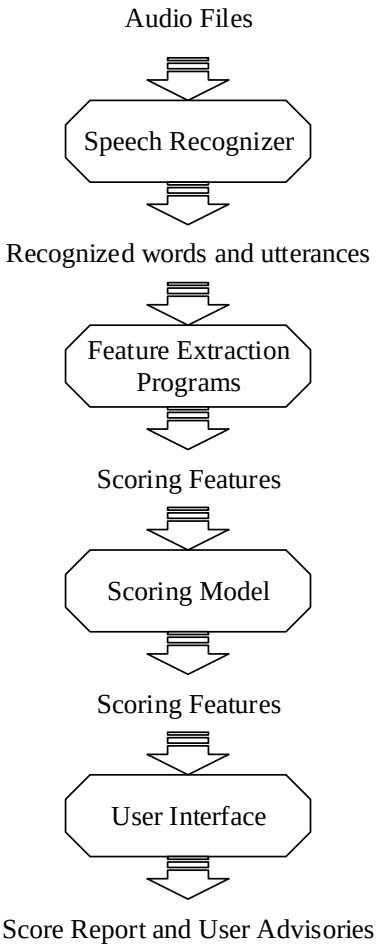


Figure 2.6. The architecture of the SpeechRater (adapted from Xi, et al., 2008, p. 19)

The TOEFL speaking rubrics, as shown in Figure 2.7, include three major dimensions of evaluation: Delivery, Language Use, and Topic Development. Delivery refers to the pace and clarity of speech. This construct further breaks down into four sub-constructs: Fluency, Intonation, Rhythm, and Pronunciation. Language Use concerns the diversity, sophistication, and precision of vocabulary use and the range, complexity, and accuracy of grammar in

speech. Finally, Topic Development refers to the discourse coherence, idea progression, and content relevance of speech (Xi, et al., 2008, p. 29).

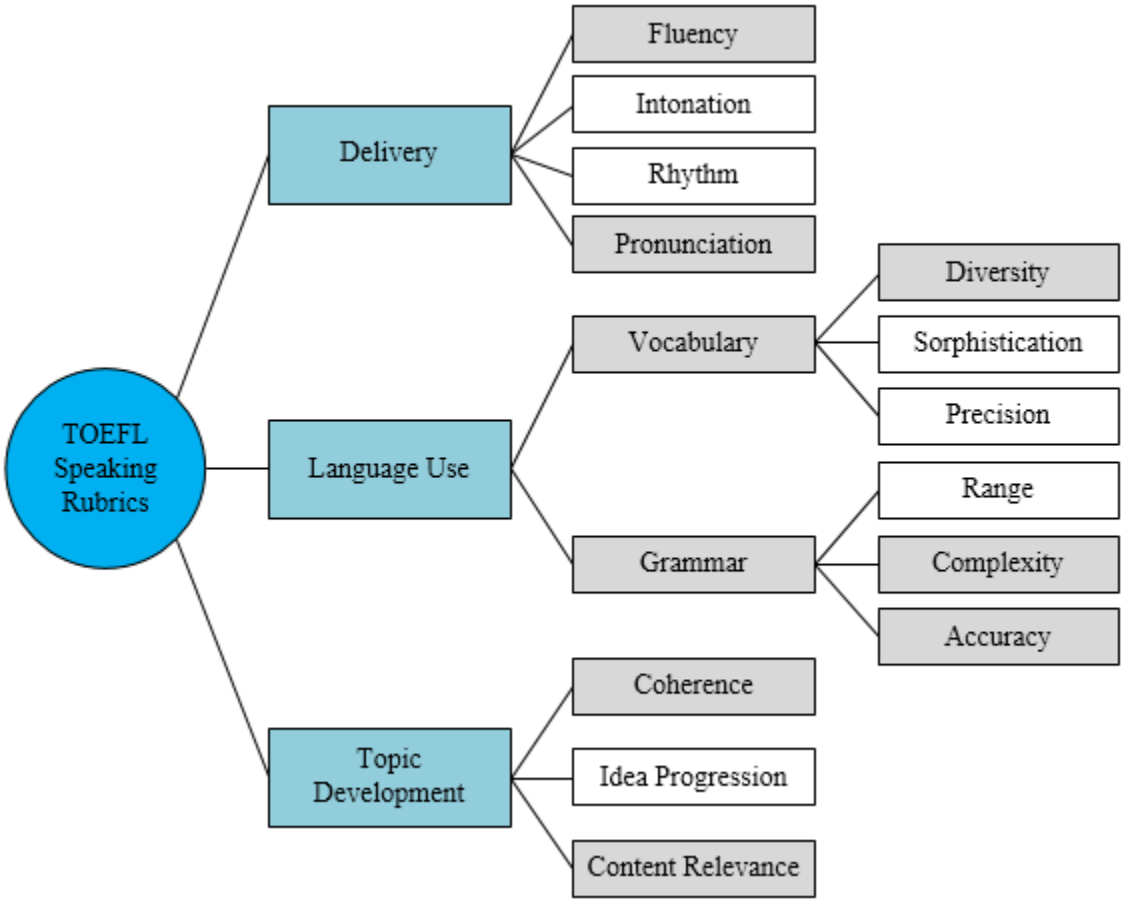


Figure 2.7. The construct coverage of SpeechRater (adapted from Xi, et al., 2008, p. 29)

The current scoring algorithms of SpeechRater, as briefly mentioned in Chapter 1, underrepresent the TOEFL speaking rubrics used by human raters (Figure 2.7) mainly because of the technological constraints in the speech recognizer. The boxes shaded in grey in Figure 2.7 represent the sub-constructs covered by SpeechRater's current Feature Extraction Programs. The generation of the scoring features in these categories, including Fluency, Pronunciation, Vocabulary Diversity, Grammar Complexity, Grammar Accuracy,

Coherence, and Content Relevance, is not significantly affected by inaccurate speech transcription. For example, the Fluency and Pronunciation features are computed based on acoustic data rather than on transcription data.

Although "understandable" and "meaningful" generation of automated scores is highly expected (Williamson, et al., 2010, p. 5), the speech features that SpeechRater can attend to at this point (Table 2.2), strictly speaking, do not meet this standard. Xi (2010, p. 293) calls them "surface linguistic features" as they are coarse and are only marginally relevant to our understanding of speaking (see a similar opinion in Weigle, 2010). For example, Fluency that is measured by average duration of speech chunks would be more interpretable if these speech chunks without exception contained coherent intonation units; Vocabulary Diversity as measured by the number of unique words in speech would be more relevant to the construct of speaking if the misused words that seriously obscure the meaning of speech and the words and phrases that are directly borrowed from the prompt were excluded. Because of these limitations, Xi addressed the strong need for adding more sophisticated scoring features into SpeechRater. In her formulation,

> The current SpeechRater model uses features that represent only a subset of the criteria evaluated by human raters and its prediction accuracy is adequate for low-stakes practice purposes but not for high-stakes decisions … It is conceivable that more high-level scoring features could be added to expand the construct coverage and improve the accuracy (Xi, 2010, p. 294).

Table 2.2 Some Representative Candidate Features of SpeechRater

Categories	Speech Features	Source
Fluency	Average duration of speech chunks, speech articulation rate, duration of silences normalized by response length in words	Zechner, Higgins, Xi, and Williamson (2009)
Pronunciation	Acoustic model score (sequence of phonemes)	Zechner et al. (2009)
Vocabulary Diversity	Unique words normalized by total word	Zechner et al. (2009)
Grammatical Complexity	The cosine similarity of part-of-speech tag distribution between the response and criterion speech samples	Yoon & Bhat (2012)
Grammatical Accuracy	Language model score (sequence of words)	Zechner et al. (2009)
Coherence	Modeling human annotation of discourse coherence (in progress)	Wang et al. (2013)
Content Relevance	The cosine similarity between the spoken response and the prompt materials	Evanini, Xie, and Zechner (2013)
	The number of word tokens that occur in both the spoken response and the prompt materials divided by the number of word tokens in the response	Evanini et al. (2013)

In addition to scoring feature development, SpeechRater developers have attempted different ways of constructing the scoring model. Xi and her colleagues (Xi, et al., 2012) compared two alternative scoring methods, multiple regression and classification trees, and concluded that a multiple regression model with regression coefficients determined by

content experts was more parsimonious and stable than a classification tree model for predicting human criterion scores of oral proficiency.

2.5.3 The Role of Construct Theory in Developing SpeechRater

SpeechRater's construct expansion is not only held back by technological constraints but also by an inadequate understanding of the L2 oral construct (Chapelle & Chung, 2010; Xi, 2008; Xi, et al., 2008). A computer has to be given explicit instructions on what features to extract from the spoken data, how to evaluate these features, and how to combine the feature scores into a holistic test score that reflects L2 oral proficiency. The creation of these scoring rules is high-stakes (Xi, 2012) and must be guided by a precise, uncontroversial definition of the test construct. Unfortunately, language testing researchers have not expressed full confidence in providing such a definition.

Some leading researchers in language testing have indicated the difficulty in reaching a consensus on the meaning of language proficiency in that various perspectives and positions can be taken in defining the construct (e.g., Bachman, 2006, 2007; Chapelle, 1998). The lack of a single best way to define language proficiency, according to Chapelle, Enright, and Jamieson (2010, p. 12), has given "enormous burden" to test validation because test construct plays a central role in Messick's (1989) validity theory.

Recently, Kane's (1992, 2006, 2013) argument-based validation framework, by unfolding Messick's (1989) meaning-overladen construct validity into to a chain of interpretive and use argument (IUA), may have downplayed the role of a theoretical construct in validation practices (Chapelle, et al., 2010; Xu, forthcoming). However, stakeholders' strong need for ASE to generate learning-oriented feature scores in addition to

a holistic score seems to have reiterated the importance of construct theory to test validation. That is, for each feature score to be interpretable and useful to language learning, their logical connections or contribution to the L2 oral construct must be explicitly articulated based on a plausible construct theory (e.g., the relationship between collocation and oral proficiency). As such, it is foreseeable that validation of ASE will stimulate concentrated research that aims to gain a better understanding of the L2 oral construct. This, in my opinion, echoes the constant theme of test validation: "Essentially, in the studies of … validity, we are validating the theory …" (APA, 1954, p. 14).

2.6 Technology of Automated Collocation Extraction and Evaluation

As mentioned in Section 2.4, constructed collocational behaviors in spontaneous learner speech assumingly contains rich information regarding the speaker's L2 oral proficiency. The question is then whether a computer can be trained to recognize the collocation occurrences in learner language. Very fortunately, this topic has been extensively researched and the results are fruitful.

In the last two decades, computational linguists have experimented with different approaches to extracting collocations from text corpora using natural language processing (NLP) technologies. It was found that symbolic systems which were built upon large lexicons and rule bases did well in identifying low-frequency collocations in texts (Michou & Seretan, 2009; Wehrli, Seretan, Nerima, & Russo, 2009); in contrast, statistically driven systems based on frequency and word co-occurrence models were efficient in extracting high-frequency collocations (Evert, 2005); further, a hybrid design by combining the two systems

into one, seemed most promising, resulting in best extraction precision and coverage (Piao, Rayson, Archer, & McEnery, 2005).

Existing collocation extraction systems, such as Collocation Tool mentioned below, tend to take a hybrid approach in virtue of the advances in parsing technologies. According to Seretan and Wehrli (2006), syntactic analysis of the source corpora has become an inescapable precondition for collocation extraction. The hybrid approach usually consists of two steps: 1) identifying candidates based on morphologic and syntactic pre-processing of the source texts and 2) ranking the collocation candidates according to collocational strength scores computed using association measures (Futagi, Deane, Chodorow, & Tetreault, 2008).

Collocation Tool is a hybrid collocation extraction program developed by ETS to detect collocation errors in ESL essay writing. By using a part-of-speech (POS) tagger, the program can identify collocation candidates that match seven target syntactic patterns, namely adjective-noun (ADJN), noun-noun (NN), noun of noun (N-of-N), verb-object (VN), adverb-verb (ADVV), verb-adverb (ADVV), and phrasal verbs (PHV). Based on frequency and rank-ratio statistics derived from two reference corpora, the program makes binary judgments (OK or ERROR) on collocation accuracy. It was found that the program had "near-human" performance in recognizing valid collocations in ESL writing (.80-.89); however, its precision in detecting collocation errors was inadequate (.28-.30); Futagi, et al., 2008).

Xu and Xi (2010) tested Collocation Tool's robustness to transcriptions of ESL spoken data based on 556 TOEFL iBT Speaking Practice Test responses. They had the same finding as Futagi et al. (2008) that Collocation Tool had high precision in identifying valid collocations (.947-.976) but low precision in detecting deviant collocations (.093-.178).

Using human judgment as the gold standard, they also found that Collocation Tool missed 37% of the collocation occurrences in the transcriptions. However, they anticipated that Collocation Tool's accuracy and coverage would improve remarkably if its reference corpora contained spoken texts rather than written texts.

2.7 Chapter Summary

This chapter carried out a literature review on the definitions of collocation, the relationship between collocational knowledge and overall L2 proficiency, the contemporary speech-processing theories, the design of SpeechRater, and NLP technologies on collocation extraction and evaluation. The main points of the chapter are summarized below:

a) Collocation, which falls into two common categories (lexical and grammatical), concerns habitual (rather than fixed) word co-occurrence, unitary processing of language, and language use in a somewhat figurative sense (Section 2.1).

b) Although there is considerable evidence for a positive relationship between L2 collocational knowledge and overall L2 proficiency, the relationship between collocation usage in naturally occurring speech and perceived L2 oral proficiency is under-researched (Section 2.2).

c) A number of contemporary speech-processing theories assume the importance of formulaic language (including collocation) to L1 and L2 speaking. It is plausible that effective use of collocations contributes to multiple aspects of oral language performance (Section 2.3).

d) It is argued that collocation is the key to assessing L2 oral proficiency; in addition, the construct of collocational knowledge assessed by collocation tests is

different from that of spoken collocational competence (SCC) that assumingly accounts for one's collocational performance in spontaneous speaking (Section 2.4).

e) Despite the constraints from speech processing technologies, SpeechRater is promising in that it has the potential for analyzing learner language in meticulous detail. The validity of the interpretation and use of automated scores largely rests upon a scientific evaluation of the construct coverage and relevance of the scoring algorithms. This evaluation has to be informed by a precise construct definition of L2 oral proficiency which seems unavailable at this point (Section 2.5).

f) It seems possible to rely on NLP technologies to identify and evaluate collocations in learner language. The Colocation Tool developed by ETS offers a good example (Section 2.6).

2.8 Research Questions

Although theory suggests a relationship between collocation use in spontaneous speaking and perceived oral proficiency, it is unclear how strong this relationship is and whether this relationship would vary in different speaking contexts. On the other hand, the research on measuring L2 learners' collocation production in naturally occurring speech is relatively scarce. It thus remains an open question as to what observed collocation characteristics in learner speech contribute to their oral performance. It was hoped that this dissertation study would fill these research gaps and set a good example of developing high-level scoring features for automated speech evaluation based on a construct theory of L2 speaking. The study posed four research questions as follows:

First, what are the basic characteristics of the collocations that ESL learners produce in naturally occurring speech?

Second, what collocation measures can effectively differentiate among ESL speakers at different oral proficiency levels?

Third, to what extent can collocation measurement predict human judgement of overall L2 oral proficiency?

Fourth, would the magnitude of the relationship between collocation measurement and human judgement of L2 oral proficiency vary across two distinct speaking contexts? In this study, the two contexts investigated were interactive daily conversation and solo presentation on an academic topic.

CHAPTER 3: METHODOLOGY

This chapter presents the methodology of this dissertation. It begins with an overview of the research design and then elaborates the procedures used for selecting ESL participants from an existing oral English exam database, transcribing the participants' test responses, and identifying and coding the collocations in the transcriptions. The chapter ends with an elaboration on the data analysis techniques employed to answer the four research questions raised earlier.

3.1 Overall Research Design

This study employed a quantitative research design to investigate the empirical relationship between ESL learners' collocation use in two communication-oriented oral English exams and their resultant oral English proficiency judged by trained human raters. In designing the study, I took a postpositivist perspective (Creswell, 2009), treating the construct of oral proficiency as a human conjecture and the investigation as a scientific inquiry of the meaning of the construct and the variation of this meaning across different contexts of language use.

The study was carried out in three phases: data collection, data annotation, and data analysis (Figure 3.1). In the phase of data collection, the participants of the study were selected and their spoken responses in the two exams were transcribed. In the phase of data annotation, the collocation occurrences in the transcriptions were manually identified and coded and a series of theory-driven collocation measures were generated based on the coding. In the last phase of data management and analysis, statistical analysis was performed on the

collocation measures and human criterion measures of oral proficiency to answer the four research questions raised in Chapter 2.

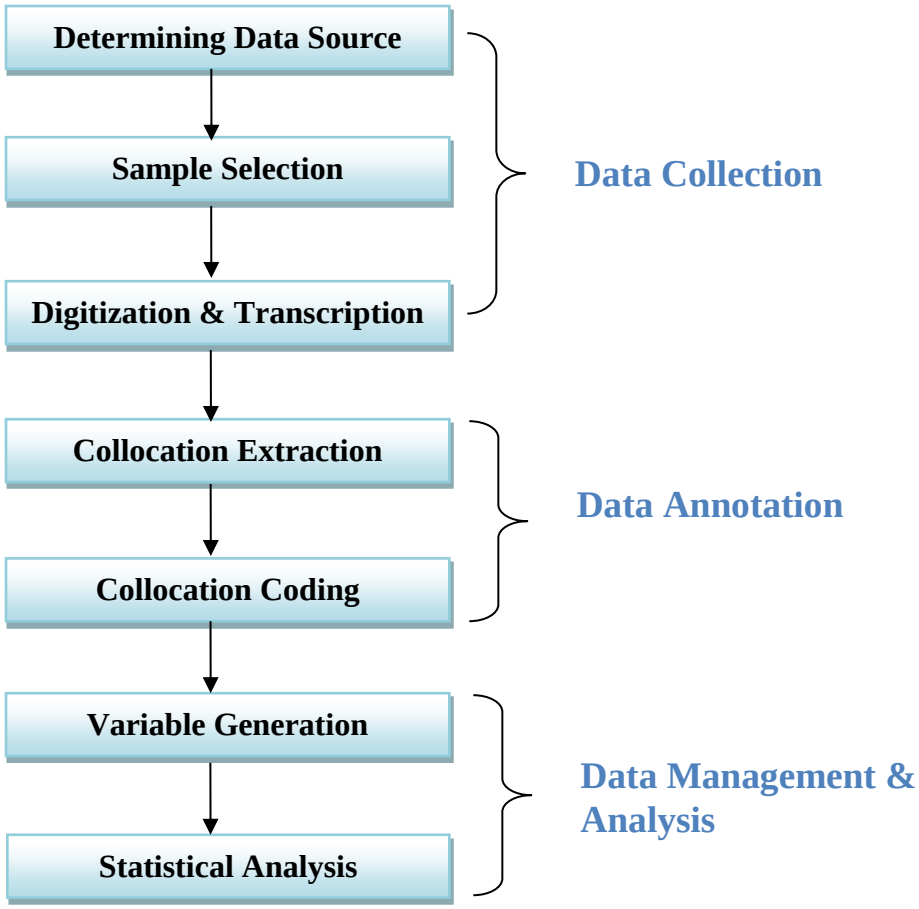


Figure 3.1. An illustration of the three phases of the dissertation study

3.2 Data Collection

The first phase of the study concerned determining the data source, selecting speech samples, digitizing spoken data, and transcribing the spoken data. This phase of work was completed between summer 2010 and summer 2013.

3.2.1 Data Source

The spoken data of this study were secondary data obtained from the SPEAK and TEACH exams at Iowa State University (ISU), which is an oral English proficiency test used to assess international teaching assistants' (ITAs) readiness for teaching. This test was chosen based on four considerations. First, the speech samples elicited by the exams were considerably long. Longer speech samples assumingly would elicit more collocation observations in L2 speech and thus allow a more reliable measurement of a speaker's oral collocation production. If the number of collocation observations is small, the random error inherent in a collocation measure will be large. This can lower the precision in estimating the true magnitude of the relationship between observed collocation performance and perceived L2 oral proficiency (see Thorndike, 1951). Second, ESL speakers of varied oral proficiency levels could be found in the exam database. Third, each examinee was required to speak in two distinctive domains of language use and their language performance in each domain was evaluated by the same raters. Fourth, the testing program was adapted from the retired Test of Spoken English (TSE) and thus adopted nearly the same scoring scale and rubrics as TOEFL iBT. The permission for using this data source for conducting this dissertation study is attached in Appendix A.

SPEAK and TEACH refer to the SPEAK interview exam and the TEACH simulation exam. SPEAK assesses the oral language ability to converse with an interlocutor on daily, campus-life topics. The exam is administered by an interviewer and contains five tasks: warm up, impromptu speaking, object description, and role-play. In contrast, TEACH assesses the language ability to disseminate academic information. The exam includes a five-minute solo presentation and a three-minute question and answer session. The examinee

chooses what to teach from a list of topics in his or her field of study and is given one hour to prepare for the presentation. To facilitate test preparation, textbook pages on the selected topic are provided. An overview of the structure of SPEAK and TEACH is illustrated in Table 3.1. The reported correlations of SPEAK and TEACH with TOEFL iBT Speaking section were .57 and .64 (International TA Program, 2012).

Table 3.1 An Overview of the Structure of the SPEAK and TEACH Exams

	Tasks	**Task description**	**Preparation**	**Duration**
SPEAK	Warm-up	Informal conversation	None	1 minute
	Impromptu speaking	Two questions with follow-up questions	None	2 minutes per question
	Object description	Describing an object (e.g., appearance, make, and function)	None	1 minute
	Role-play	A conversation for a specific communicative objective	1 minute	1-2 minutes
TEACH	Mini-lecture	Solo presentation	1 hour	5 minutes
	Question handling	Responding to questions	None	3 minutes

Both SPEAK and TEACH are holistically scored on a scale of 0 to 300 (with increments of ten), thus equivalent to the scoring scale of 0-30 used by TOEFL iBT Speaking section. This scoring scale is further divided into seven score bands: not competent (0-110), not adequate (120-160), very limited (170-190), limited (200-210), adequate (220-240), strong (250-270), and excellent (280-300). Detailed descriptors of each score band are provided for rater training and score interpretation. These descriptors cover a wide range of spoken language qualities and features, including language competencies (functional,

strategic, linguistic, sociolinguistic and discourse), pronunciation, vocabulary and grammar, and pace and fluency.

The SPEAK and TEACH exams are scored by a panel of three raters live. These scores are averaged and rounded to obtain the final test score. However, if two raters disagree beyond 30 points, a fourth rater is asked to score the video or audio recording of the test. The most deviant score among the four is discarded before calculating the final score. The testing program delivers regular rater training sessions (before each testing cycle and each summer) and rater certification assessments to calibrate raters and minimize rater bias in rating and interacting with the examinees (see e.g., Brown, 2012; McNamara, 1996). The reported inter-rater reliabilities based on intraclass correlation of three concurrent ratings were .89 for SPEAK and .91 for TEACH (International TA Program, 2012).

Based on the test results, the testing program places examinees into four levels of oral proficiency and makes recommendations about their suitable teaching duties. Level 1 examinees (who score 220 or higher in both exams) are fully certified. They are granted the permission to perform most teaching assignments on campus such as giving a lecture, leading group discussion, and tutoring students in a help room. Level 2 examinees (who score 220 or higher in one exam and between 200 and 210 in the other) are certified with a minor restriction. That is, they are not encouraged to teach a class on their own. Level 3 examinees (who score between 170 and 210 in both exams) are certified with more restrictions. It is suggested that their teaching assignments only require a limited amount of oral communication (e.g., being a lab monitor). Level 4 examinees (who score 160 or below in one of the exams) are not certified. They are only eligible for carrying out jobs that do not involve oral communication such as grading, setting up lab equipment, and maintaining a

course website. Generally, ITAs who are placed in Level 1 and Level 2 are considered as accomplished ESL speakers. Those in Level 3 and Level 4 are below the cut-off criteria; they are required to take some remedial courses on oral English communication skills before taking the exams again.

3.2.2 Sample Selection

The participants of this study were sixty Chinese-speaking ITAs chosen from the SPEAK and TEACH exam database. The target population of this study was Chinese-speaking ESL graduate students newly admitted at Iowa State University. The sampled population was 356 Chinese-speaking ITAs who took the SPEAK and TEACH exams for the first time between 2006 and 2011 and who gave written informed consent for their test responses and results released for L2 assessment research. It was assumed that the sampled population and the nonsampled population were generally similar in terms of their collocational behaviors in English speaking.

The study employed a single-stage stratified sampling procedure to select the sample. The examinees' test identification numbers were used to construct the sampling frame. The sampling frame was first divided into four strata which were defined as the four levels of oral proficiency that ITAs were placed into as a result of the SPEAK and TEACH exams. Then, fifteen different examinees were randomly selected from each stratum without replacement to obtain the sixty-participant sample. In the sampling process, no examinee identification information except test identification number, native language, proficiency level, and exam scores was disclosed.

ITAs' native language was controlled in this study based on three considerations. First, previous research has indicated that cross-lingual distance (i.e., the similarity between L1 and L2) is a confounding factor that may mediate the relationship between L2 proficiency and L2 collocational performance (e.g., Nesselhauf, 2005). For this reason, L2 collocation research usually focuses on a single L1 group (e.g., Laufer & Waldman, 2011; Nesselhauf, 2005; Voss, 2012, Wood, 2007). The present study also followed suit to eliminate this confounding factor. Second, nearly half of the examinees in the SPEAK and TEACH exam database (48%) were native speakers of Mandarin Chinese. Considering that Chinese students comprise the largest international student body in the United States (Institute of International Education, 2012) and the biggest population of TOEFL iBT test takers (Powell, 2012), it became a priority to investigate this particular L1 group. Third, the researcher was a native speaker of Mandarin Chinese. Speaking the same L1 as the research subjects gave him a great advantage in speech transcription, collocation coding, and interpretation of the findings.

3.2.3 Digitization and Transcription

The sixty Chinese-speaking ITAs' test recordings were first digitized. The exams administered before 2009 were recorded on videotapes. They were first converted into MP4 video files using a Panasonic DVD/VCR player and a Pinnacle video transfer device. Afterwards, all the digital video files, including the exams administered after 2009, were converted into Windows Media Audio files using a software program called AimOne Video Converter. Audio information that revealed an examinee's name was removed.

The speech samples of this study were collected from the sixty participants' warm-up, impromptu speaking, and object description tasks in the SPEAK exam and mini-lecture in the TEACH exam. Their spoken data generated in the role-play task of the SPEAK exam and in the question-handling task of the TEACH exam were not collected. Some of these data were missing, badly recorded, or incomplete.

The speech samples were first transcribed by three trained native-English-speaking assistants and then proofread by the researcher. Nonexistent words in L2 speech were spelled out in words rather than phonetic symbols. The interviewer's speech was italicized and kept in square brackets. Disfluency features and unintelligible speech were annotated based on a transcription scheme adapted from Du Bois (1991) (Table 3.2). Through proofreading, the researcher managed to clear most cases of transcribers' uncertain hearing. However, a very small amount of unintelligible speech remained in the TEACH exam transcriptions. All transcriptions were saved as Microsoft Word documents.

Table 3.2 Transcription Scheme Adapted from Du Bois (1991)

Speech Features	**Definition**	**Symbol**	**Example**
Unacceptable pause	Discontinuity in speech that is noticeable to the transcriber	Three dots …	I don't … normally sound like that.
False start	A speaker utters the initial portion of speech but abandons it before finishing.	Single hyphen -	I don't- but I have a cold today.
Restart or self-correction	A speaker repeats or rephrases a small portion of previous speech.	Single hyphen -	I don't- I don't normally sound like that.
Uncertain hearing	Portions of speech that is not clearly audible to the transcriber.	A pair of angle brackets < >	I don't <normally> sound like that.
Indecipherable word or syllable	Words or syllables which are not audible enough to allow a reasonable guess	A capital letter X	I don't X sound like that.

3.3 Data Annotation

In the second phase of the study, the collocations in the transcriptions were manually identified and coded by human experts. A pilot study on twenty speakers' SPEAK responses was conducted in the Spring 2014 semester. A full-scale study based on sixty speakers' SPEAK and TEACH responses was performed in the Fall 2014 semester and the Spring 2015 semester with the support of the TOEFL Small Grants for Doctoral Research.

3.3.1 Target Collocations

This study chose to only investigate lexical collocations due to their importance to speech comprehensibility. Lexical deviation or *misselection*, according to some linguists, is more likely to obscure meaning than structural deviation or *misformation* (Barnbrook, 2007; Bolinger, 1976; Hunston, 2002; Sinclair, 1991). In the phrase "severely [seriously] interested on [in] the job" for example, it is probably hearing the deviant adverb "severely" than the improper preposition "on" that surprises an addressee more. There is already empirical evidence to show that lexical deviation causes more processing difficulty than structural deviation to native speakers (Millar, 2011, p. 141). Moreover, a speaker tends to pronounce content words more prominently than grammatical words in spoken English (Celce-Murcia, Brinton, & Goodwin, 1996). Hence, lexical deviation may be more noticeable than grammatical deviation.

Grammatical collocation was excluded also in consideration of feasibility. Recently, the definition of grammatical collocation has been expanded from a lexical-grammatical combination such as "interested in" (Benson, et al., 1986; Benson, Benson, & Ilson, 1997) to an entire syntactic environment of a lexical item such as "interested in doing" (Gries, 2008;

Hunston, 2002; Hunston & Francis, 2000; Stefanowitsch & Gries, 2003). Under the latter broader definition, almost everything in language seems to belong to a category of grammatical collocation. Then, setting clear-cut rules for identifying grammatical collocations in L2 speech would be extremely demanding, not to mention that syntactic variations around lexical items are common in oral language and that L2 learners' syntactic constructions based on L2 declarative rules are rather creative and unpredictable.

Following Nesselhauf's (2005) phraseological approach, this study focused on ten syntactic patterns of lexical collocations: 1) adjective and noun, 2) adverb and adjective, 3) adverb and verb, 4) noun and noun, 5) noun of noun, 6) noun and verb, 7) verb and noun, 8) phrasal verb and adverb, 9) noun and phrasal verb, and 10) phrasal verb and noun. Examples from the real data of this study are provided in Table 3.3.

Among them, the first seven patterns are traditional, which can be found in Benson, et al.'s (1997) *BBI dictionary of English word combinations*. They are also targeted by ETS's Collocation Tool (Chapter 2, Section 2.6). The last three which contain a phrasal verb are new additions. The phrasal verbs in English often convey idiomatic meanings. For this reason, linguists believe that they pose more problems for ESL learners than single-word verbs (Biber, et al., 2002; Sinclair & Moon, 1989).

Table 3.3 Ten Target Collocation Patterns

	Syntactic Patterns	Example (Collocation ID)	Exam
1	Adjective and noun (ADJ-N)	social network (111021303) logical connection (120717413)	SPEAK TEACH
2	Adverb and adjective (ADV-ADJ)	equally important (210703718) closely related (320722406)	SPEAK TEACH
3	Adverb and verb (ADV-V)	privately owned (111021323) randomly selected (220718917)	SPEAK TEACH
4	Noun and noun (N-N)	college education (110710922) energy intake (121100134)	SPEAK TEACH
5	Noun of noun (N-of-N)	tail of a kite (110720611) steepness of a line (120917903)	SPEAK TEACH
6	Noun and verb (N-V)	a plane lands (310913805) demand increases (420710506)	SPEAK TEACH
7	Verb and noun (V-N)	fly a kite (110720612) consume oxygen (121100127)	SPEAK TEACH
8	Phrasal verb and adverb (PHV-ADV)	get up early (311019901) N/A	SPEAK TEACH
9	Noun and phrasal verb (N-PHV)	dreams come true (311017210) starch breaks down (220901109)	SPEAK TEACH
10	Phrasal verb and noun (PHV-N)	put out a fire (410701208) cut down cost (321017521)	SPEAK TEACH

3.3.2 Collocation Extraction and Validation

The researcher extracted the collocations from the transcriptions of L2 speech based on the following criteria:

a. A collocation string must match one of the ten target syntactic patterns.

b. The two content words in a collocation string are semi-fixed. That is, they are substitutable, however, with certain restrictions.

c. If a collocation string contains a pronoun that refers to a specific regular noun that appears earlier, the antecedent of the pronoun is also extracted and placed in a parenthesis. For example: look it up (a dictionary).

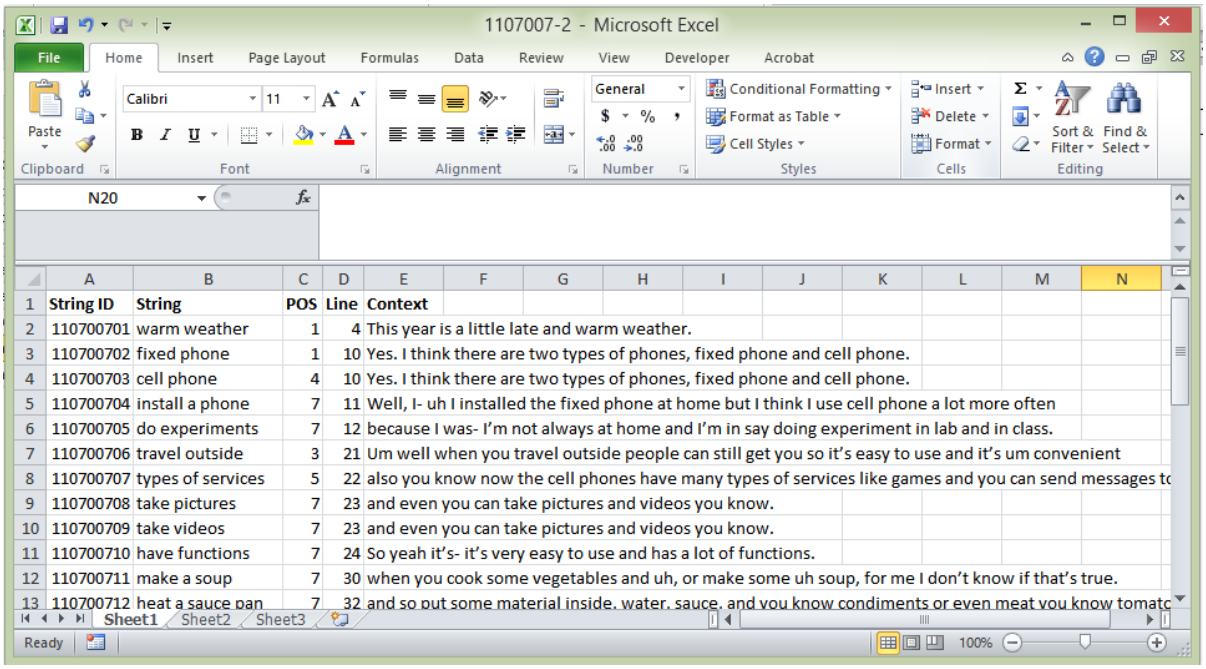
d. A recurrent collocation is only extracted once when it appears for the first time. Repeated vocabulary use only indicates the speaker's limited vocabulary size (Xi, et al., 2008, p. 35). Overusing certain words or phrases would not inflate the perceived oral proficiency according to the TOEFL scoring rubrics (see 'vocabulary diversity' in Chapter 2, Section 2.5.2).

e. A collocation string borrowed from the interviewer' speech or the prompt is excluded. In conversations, it is common that two interlocutors make use of each other's choices of words (Garold & Pickering, 2004). However, in the context of oral language assessment, the behavior of borrowing linguistic resources from a more advanced speaker, in my opinion, cannot be deemed as evidence for proficiency. Borrowing may be a way of learning or simply a strategy that L2 learners adopt to fill their lexical gaps (see Bygate, 1987).

f. Proper nouns (e.g., Los Angeles Lakers, TOYOTA dealer) and technical terminology (e.g., 'shear lag' from Aerospace Engineering, 'tonal values' from Architecture) are excluded. Proper nouns are, strictly speaking, not collocations but fixed expressions referring to specific items, organizations, or places. On the other hand, technical terminology is collocation specific to an academic discipline and is thus incomprehensible to naïve listeners. The use of technical terminology may indicate a speaker's oral proficiency in a narrow language use domain.

Unfortunately, the collocation coders of this study were a group of naïve listeners who were unable to judge the acceptability of this type of collocations.

The manual collocation extraction consisted of three steps: marking, recording, and reviewing. The researcher first went through the transcriptions and highlighted the collocation strings that met the above criteria. Then, he created Excel spreadsheets to record these collocations. Each spreadsheet contained the collocation strings, their identification numbers, their syntactic or part-of-speech (POS) coding, their locations (Line) in the transcription, and the context from where these strings had been extracted (Figure 3.2). Finally, the researcher reviewed the extraction with the same criteria. These spreadsheets were later sent to human coders for collocation coding.



	A	B	C	D	E
1	String ID	String	POS	Line	Context
2	110700701	warm weather	1	4	This year is a little late and warm weather.
3	110700702	fixed phone	1	10	Yes. I think there are two types of phones, fixed phone and cell phone.
4	110700703	cell phone	4	10	Yes. I think there are two types of phones, fixed phone and cell phone.
5	110700704	install a phone	7	11	Well, I- uh I installed the fixed phone at home but I think I use cell phone a lot more often
6	110700705	do experiments	7	12	because I was- I'm not always at home and I'm in say doing experiment in lab and in class.
7	110700706	travel outside	3	21	Um well when you travel outside people can still get you so it's easy to use and it's um convenient
8	110700707	types of services	5	22	also you know now the cell phones have many types of services like games and you can send messages to
9	110700708	take pictures	7	23	and even you can take pictures and videos you know.
10	110700709	take videos	7	23	and even you can take pictures and videos you know.
11	110700710	have functions	7	24	So yeah it's- it's very easy to use and has a lot of functions.
12	110700711	make a soup	7	30	when you cook some vegetables and uh, or make some uh soup, for me I don't know if that's true.
13	110700712	heat a sauce pan	7	32	and so put some material inside, water, sauce, and you know condiments or even meat you know tomato

Figure 3.2. The screenshot of a collocation coding spreadsheet

To check the accuracy of his extraction, the researcher had a second coder, a native Chinese speaker (who holds a master degree in ESL) do the same job with the pilot study sample of twenty speakers. The second coder independently extracted 392 collocations; she then she compared her extractions with the 412 collocations that the researcher had identified earlier and marked the discrepancies. After that, they adjudicated on the discrepancies through discussion. The 391 collocations agreed upon by both coders were considered as the gold standard or 'true collocations' based on which, two statistics, precision and recall, were computed (see Section 3.3.1 for more detail). As the two statistics turned out to be satisfactory, the researcher performed collocation extraction on the remaining spoken data of forty speakers on his own.

3.3.3 Collocation Coding Schemes

The development of collocation coding schemes was probably the most important step in this dissertation study. A review of the literature suggests that an L2 learner's collocation usage in spontaneous speech contains rich information about his or her oral language ability (Chapter 2, Sections 2.3.5. and 2.4). Following Ellis and Barkhuizen's (2005) approach to analyzing learner language, the researcher decided to evaluate three dimensions of oral collocational performance: accuracy, complexity, and fluency (Figure 3.3). The first two dimensions focus on the lexico-grammatical characteristics of collocation and correspond to 'vocabulary precision' and 'vocabulary sophistication' in the TOEFL speaking rubrics. The third dimension concerns phonological coherence or smoothness of collocation articulation; it corresponds to 'rhythm' in the TOEFL speaking rubrics (see Chapter 2, Section 2.5.2). The three dimensions of collocational performance are assumingly

interrelated because they all draw on and compete for the limited attentional resources. Based on the trade-off hypothesis (Skehan, 1998, 2009), a speaker may sacrifice one aspect of language performance for another in on-line speech production. To measure the various qualities in learner collocational performance, subcategories of each dimension were created and the corresponding coding schemes were developed.

The dimension of accuracy was divided into two subcategories: *semantic accuracy* and *grammatical accuracy*. Semantic accuracy refers to the meaningfulness of the co-occurrence of two content words (lemmas) in the speaking context regardless of possible morpho-syntactic errors. The coding scheme of semantic accuracy contained three levels: unacceptable, substandard, and acceptable (Table 3.4). In contrast, grammatical accuracy concerns the minor surface errors in the collocation string, such as determiner errors, wrong word forms, and function word (e.g., preposition) errors (Table 3.5). The coding scheme on grammatical accuracy was dichotomous: error-free and erroneous. The dual layer of accuracy coding was informed by Levelt's (1989, 1999a, 1999b) conceptualization of speech encoding. Levelt proposes that *lemma selection* preludes *syntactic composition* and either process may introduce errors into the end product of overt speech (see Chapter 2, Section 2.3.2 above).

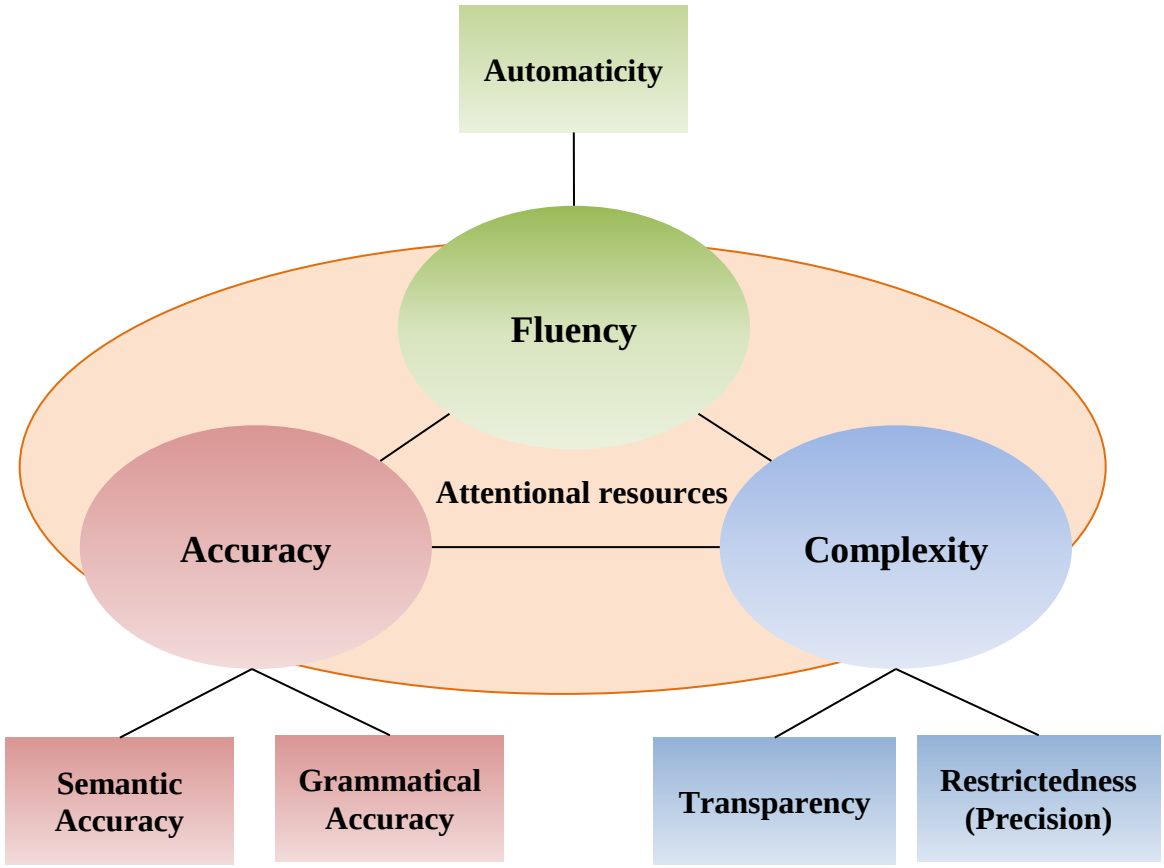


Figure 3.3. Three dimensions of collocation measurement

Table 3.4 The Coding Scheme of Semantic Accuracy

Levels	Examples from Real Data
Unacceptable A collocation is nonsense in the speaking context. Using the collocation obscures the meaning of the entire sentence.	[Where have you gone to look for furniture?] Um, resort to some friends of mine or just pick up from the street (110717402). Correction: ask friends?
Substandard The meaning of a collocation is understandable. However, the collocation sounds awkward in the speaking context.	… like if you would like to be plumber, you don't have to go to college, you go to training school and then they will be fine (110710921). Correction: vocational school
Acceptable The collocation makes perfect sense in the speaking context.	… and you must practice a lot and from a coach, help you get through the road test (110916715).

Table 3.5 A Summary of Surface Error Types

Types of Errors	Description	Examples from Real Data
Determiner errors	A determiner (a, an, the, this, those, his, her, some, any, few, the same, etc.) is omitted, extraneous, or used improperly.	And uh, yeah probably you can open online store, and try to invest in the stock or something (111017117). Correction: a store
Wrong word forms	Derivational or inflectional errors (singular/plural, tense, etc.)	And secondly uh I think the government and the industry should uh should- should achieve agreement on- on some regulations uh on how to- how to control um those- those air pollutioners from the industry (110717423). Correction: air polluters
Function word errors	A function word (prepositions, words of degree, particles, etc.) is wrong.	… then I uh check on my email and uh maybe I, I just chat with my friends back in China a little (210801005). Correction: check email

The dimension of complexity contained two subcategories: *transparency* and *restrictedness*. Transparency refers to the degree to which the meaning of a collocation is literal. The coding scheme of transparency had two levels: partially figurative and literal (Table 3.6)—fully figurative expressions are idioms rather than collocations (see the definition of collocation in Chapter 2, Section 2.1). Research has indicated that the acquisition of idiomatic expressions in L1 and L2 is difficult and requires a great amount of exposure to the target language (Bortfeld & Brennan, 1997; Crutchley, 2007; Skripnikova, 2012). This suggests that collocation figurativeness may be an indicator of vocabulary sophistication. Restrictedness, on the other hand, concerns the substitutability of the modifier (i.e., the subordinate word) in a collocation. A modifier is the element other than the headword in a collocation. For example, in the collocation ‘deeply involved’, the word ‘deeply’ is a modifier that imparts a sense of degree to the headword ‘involved’. The coding

scheme of restrictedness included two levels: highly restricted and moderately restricted—unrestricted modifiers are not found in collocations but in free combinations which were not considered in this study (e.g., a good teacher). For a highly restricted collocation, it is almost impossible to find a replacement for its modifier while keeping its meaning intact. Stated another way, highly restricted collocation indicates extremely precise, irreplaceable vocabulary use (Table 3.7). It is notable that only acceptable collocations (the highest level in semantic accuracy) were coded for complexity. When the meaning of a collocation is ambiguous, it is difficult to make judgment on its transparency and restrictedness.

Finally, the dimension of fluency had a single subcategory: automaticity. Automaticity was defined as the coherence or smoothness of collocation utterance. The coding scheme on automaticity had two levels: uninterrupted utterance and interrupted utterance. Uninterrupted utterance means that the speech context of a collocation (a range between three words before the first content word and three words after the second content word *in the transcription*) does not contain any disfluency feature. In other words, the collocation is articulated automatically as a "unitary, unanalyzed phrase" (van Lancker, et al., 1981, p. 330). On the contrary, interrupted utterance means that the speech context of a collocation is contaminated by disfluency features such as restart, self-correction, hesitation (unacceptable pause or filler), and repetition. The observed disfluency nearby or within a collocation utterance may imply that the speaker allocates a large amount of attentional resources for lexicalizing the conceptual chunk (by searching the lexicon, drawing on L2 declarative rules, or applying compensatory strategies); because of the depletion of attentional resources, the speaker temporarily loses control of speech formulation (Kormos, 2006; Levelt, 1999a, 1999b), resulting in a slip in the overt speech.

Table 3.6 The Coding Scheme for Transparency

Levels of coding	Examples from Real Data
Partially figurative A collocation contains both figurative and literal elements. For example: • crack a code • blow a chance	Like you have to hook your readers (120710915). Uh, we usually use a fire extinguisher to uh, put out the- put out the fire (410701208).
Literal All elements in a collocation are used in a literal sense. For example: • crack a nut • blow a trumpet	This weekend maybe looking for some second-hand furnitures to furnish my apartment (210715503). But in another context, senior people are not so familiar with some new things like computers, and computer games (210715522).

Table 3.7 The Coding Scheme for Restrictedness

Levels of coding	Examples from Real Data
Highly restricted It is almost impossible to find a substitute for the modifier in the collocation. For example: • *Blow* a chance • an *act* of violence • Bees *buzz*	Uh we are using the *lunar* calendar so it's different from what we are using in the U.S (111020912). I think we *brush* our teeth twice every day like in the morning and in the evening before we go to bed (210703711). So this is the back of the chair and this is *arm* support (121021416).
Moderately restricted The modifier may be substituted with a small number of words. For example: • Deeply (closely, heavily, intimately, very much) involved • Gain (acquire/obtain/advance/expand/increase) knowledge	… I think the two are the major reason (210805216). (chief, key, main, primary, principal) reason And I think that make me to be a- more like a man at that time after I married, to have the responsibility for our home and to earn more money (210901115). (take, bear, shoulder, carry, assume) responsibility

Table 3.8 The Coding Scheme for Automaticity

Levels of coding	Examples
Uninterrupted Utterance The collocation is a coherent phonological unit. That is, the utterance is smooth and automatic.	So you should pay attention to that especially during the midterm exam (220804720).
Interrupted Utterance The utterance of the collocation is interrupted by disfluency features such as self-correction, unacceptable pause or hesitation,	Ok so, first I think the government should set up a very strict rule to uh to- to- *ban* the air *pollution* (110700719).

A summary of the coded collocation features, including feature definitions, coding scale, and scale range, are provided in Table 3.9 below.

Table 3.9 A Summary of the Coded Collocation Features

Features	Definition	Scales	Scale Ranges
Semantic Accuracy	The meaningfulness of the combination of two content words in a collocation regardless of morpho-syntactic errors	Ordinal	1 (Unacceptable) 2 (Substandard) 3 (Acceptable)
Grammatical accuracy	Existence of minor surface errors such as determiner errors, wrong word forms, and function word errors	Ordinal	0 (Error-free) 1 (Erroneous)
Transparency	The literalness of the content words in a collocation	Ordinal	0 (Literal) 1 (Partially figurative)
Restrictedness	The substitutability of the modifier or subordinate word in a collocation	Ordinal	0 (Moderately restricted) 1 (Highly restricted)
Automaticity	The smoothness of collocation utterance	Ordinal	0 (Uninterrupted Utterance) 1 (Interrupted Utterance)

3.3.4 Procedure of Collocation Coding

The collocation strings extracted from the transcriptions were coded by nineteen human coders (including the researcher) for the above-mentioned five performance subcategories: semantic accuracy, grammatical accuracy, transparency, restrictedness, and automaticity. The coders consisted of four faculty members and nine doctoral students in an Applied Linguistics program, four ESL instructors holding a master degree in Applied Linguistics, and two undergraduate students majoring in linguistics. Among them, fourteen were native English speakers and five were non-native speakers. Although non-native coders were used, they were not assigned to code semantic accuracy and restrictedness because judgment on these two aspects of collocation usage requires native-speaker intuitions (Schmitt, 1998).

Coding of the entire dataset was completed in two stages. The pilot study data (twenty SPEAK responses) were coded by seven native coders. This group of coders was given training and practice (familiarization with the coding schemes, coding a small portion of data, and receiving feedback) in two face-to-face meetings with the researcher. The remaining data (forty SPEAK responses and sixty TEACH responses) were coded by nine native coders and five non-native coders. For logistic reasons, rater training and practice at this stage was given in the form of email correspondences.

In both stages, coders coded collocations via Excel spreadsheets prepared by the researcher (Figure 3.2). The spreadsheets contained isolated collocation strings extracted from the transcript as well as the contextual information (approximately one to two sentences). Coders were required to read the contextual information carefully before making their judgments. In addition to assigning codes, they made corrections and/or justified their

coding in a ‘correction/comment’ field when necessary. A total of 2172 entries of corrections or comments were provided by the coders in the spreadsheets. This information was useful for reconciling the discrepancies between two coders later. All coders worked remotely and communicated with the researcher via email for questions. The coders raised twelve questions. Among them, nine addressed specific concerns that the coders had encountered in the coding process (e.g., determining the headword of a collocation, determining the level of transparency of a light verb, coding verb-noun collocations in a passive voice); two reported technical issues (i.e., data organization and missing data); one solicited further instruction on coding collocations from an unfamiliar discipline.

Among the five subcategories, semantic accuracy, grammatical accuracy, transparency, and restrictedness were independently coded by pairs of coders whereas automaticity was coded by the researcher alone. The first four subcategories were double-coded because they involved subjective judgment on L2 collocation usage which typically provokes minor rater disagreement (Schmitt, 1998). In this study, the discrepancies between two coders were resolved by calling upon a third rater. In a small number of cases when a third rater indicated the difficulty in determining the typicality of a word combination from an academic discipline, the researcher made a final decision by consulting the Corpus of Contemporary American English (Davies, 2012), a 450 million words native-speaker corpus, and other two coders’ comments and justifications. Differently, automaticity was single-handedly coded by the researcher himself considering that the judgment on whether or not a collocational utterance contained disfluency (which had been double-annotated by a transcriber and the researcher in the transcription) was relatively straightforward and objective.

3.4 Data Management and Analysis

The human collocation coding resulted in quantitative data regarding the qualities of each collocation instance extracted from the transcriptions. A necessary step before using the data to answer the four research questions raised in Chapter 2 was to check the reliabilities of collocation extraction, collocation coding, and oral proficiency measures. As measurement experts (e.g., Haertel, 2006, Thorndike, 1951) noted, if measurement errors inherent in two variables are large, the actual magnitude of the relationships between them could be masked. On the other hand, for the collocation coding data to be useful for predicting oral proficiency, they had to be organized into operational variables that reflected each speaker's five-dimensional collocation performance in speaking, including semantic accuracy, grammatical accuracy, transparency, restrictedness, and automaticity.

To answer the four research questions raised in Chapter 2, different data analysis techniques were employed. Descriptive statistics and paired-sample Wilcoxon Rank test was used to answer the first question regarding the basic characteristics of the collocation occurrences in the two oral exams. Analysis of variance (ANOVA) was performed to answer the second research question concerning the differences of the collocation measures across proficiency level groups. Pearson and point biserial correlation and multiple and logistic regression were performed to answer the third research question on the prediction for oral proficiency based on collocation measures. Finally, a cross-validation technique based on regression was performed to answer the last question about the context effect on the oral proficiency prediction. The detailed procedure of data organization and analysis is described below.

3.4.1 Evaluation of Manual Collocation Extraction

To ensure that his manual collocation extraction was accurate, the researcher calculated precision and recall, two statistics commonly used to evaluate NLP applications (see e.g., Futagi, et al., 2008), based on the pilot study sample. Precision here refers to the proportion of true collocations (agreed upon by two coders) to all the collocations extracted by a single coder. Recall refers to the proportion of the true collocations extracted by a single coder to all of the true collocations that exist in the transcriptions. Put simply, precision concerns the accuracy of extraction whereas recall concerns the completeness of extraction. Equations (1) and (2) show how these two statistics were calculated.

$$Precision = \frac{|Collocation_{extracted} \cap Collocation_{true}|}{|Collocation_{extracted}|} \quad (1)$$

$$Recall = \frac{|Collocation_{extracted} \cap Collocation_{true}|}{|Collocation_{true}|} \quad (2)$$

3.4.2 Development of Collocation Measures

To quantify L2 speakers' collocational performance, the researcher developed eleven measures, namely Operational Collocation Performance Measures (OCPMs), based on human collocation coding. The conceptualization of these measures was inspired by the scoring features targeted by the SpeechRater (see Chapter 2, Section 2.5.2) and Hollinger's (2004) player efficiency rating (PER) of professional athletes' game productivity.

The OCPMs of interest were derived from the five subcategories of collocation coding (Table 3.10). The measures of semantic accuracy included normalized frequency[4] (i.e., per 500 words) of acceptable collocations (ACP_OK), that of unacceptable collocations (ACP_ERR), and the ratio of acceptable collocations to unacceptable collocations (ACP_RAT). The measures of grammatical accuracy were normalized frequency of error-free collocations (GRA_OK), that of erroneous collocations (GRA_ERR), and the ratio of error-free collocations to erroneous collocations (GRA_RAT). The measures of restrictedness included normalized frequency of highly restricted collocations (RES_FRE) and the proportion of highly restricted collocations to all collocation attempts (RES_PRO). The measures concerning transparency and automaticity were normalized frequency of partially figurative collocations (TRAN) and the proportion of interrupted (i.e., disfluent) collocation utterances to all collocation attempts (CHOP), respectively. It was hypothesized that advanced speakers' speech would contain more acceptable, error-free, highly restricted (precise), partially figurative, and/or phonetically coherent collocations than less proficient speakers' speech.

Additionally, a composite measure, namely Collocational Performance Rating (CPR), was created to summarize a speaker's holistic collocational performance in a speaking activity. Hollinger's (2004) PER formula sums up a basketball player's positive accomplishments (e.g., three-pointers, assists, and rebounds), subtracts the negative ones (e.g., missed shots, turnovers, and personal fouls), and weight each term based on their contributions to team performance (see also Alamar, 2013). Thus, the larger a player's PER value is, the more valuable he or she is to a team. Analogous to PER, CPR (see its formula in

[4] All frequency measures discussed below were normalized to the number of occurrences per 500 words (i.e., by dividing the raw frequency by total word count in the transcript and then multiplying 500) as 500 words were the approximate average length of each SPEAK and TEACH response.

Table 10) rewards positive collocation features, penalize negative ones, and weight each term according to their presumed effects on the general effectiveness of L2 oral communication. The heavily weighted terms (coefficient = 2) were acceptable collocations, highly restricted (precise) collocations, and phonetically interrupted collocations. The moderately weighted terms (coefficient = 1) included substandard collocations, partially figurative collocations, and unacceptable collocations. The least weighted term (coefficient = 0.5) was the collocations that contained surface errors. It was expected that CPR would demonstrate a positive association with a criterion measure of oral proficiency.

It is notable that substandard collocations were considered a positive term in the CPR formula. This is because such expressions, according to the coding scheme, do not obscure the meaning of speech. In fast-paced oral communication, an addressee may undergo many 'anticipatory processes' (Millar, 2011, p. 143), i.e., drawing on collocational knowledge from their own lexicon to rectify trivial lexical deviations (see also Pickering & Garrod, 2007). In such cases, substandard collocations, although slightly increasing the addressee's processing burden, may still contribute to speech comprehensibility.

The weight assignments in the formula were guided by two principles. First, the weights should reflect the gravity of learner errors or the order of importance of collocation features to effective oral communication. Among the negative terms, using unacceptable collocations or *lexical misselection* (coefficient = -1) was penalized more severely than using ungrammatical collocations or *misformation* (coefficient = -0.5)—as argued above (Section 3.2.1), misselection assumingly causes more ambiguity in meaning than misformation. Disfluencies or interruptions in collocation utterances were harshly punished (coefficient = -2) because they break normal speech rhythm and interfere with comprehensibility (see Brown et

al., 2005, p. 32). Among the positive terms, acceptable collocations (coefficient = 2) naturally received a heavier weight than substandard collocations (coefficient = 1). Correct use of highly restricted collocations and partially figurative collocations were also heavily rewarded (coefficient = 2) for their semantic difficulty/complexity to ESL learners (e.g., Liao & Fukuya, 2004; Nesselhauf, 2005).

Second, the scoring rule should encourage the use of sophisticated, high-quality collocations. Higgins (1997) hypothesized two types of motivation behind human behaviors: promotion focus or seeking accomplishments and prevention focus or seeking safety (see also Halvorson & Higgins, 2013). Higgins' motivation theory also applies to learner behaviors in an L2 oral proficiency exam. Examinees, on the one hand, aspire to demonstrate their language ability but, on the other, try best to hide their language deficiency (see 'compensatory strategies' in Chapter 2, Section 2.3.1). For an L2 oral exam to be informative to decision-making and instruction (Wolf, et al., 2014), its scoring rule must be designed in a way to encourage promotion-focused performance so that examinees are willing to push beyond their limits (e.g., attempting complex language and topics) and take risks of revealing their weaknesses. Based on this consideration, the positive terms regarding sophisticated collocation usage (e.g., acceptable collocations, highly restricted collocations, partially figurative collocations) in CPR were generally given a heavier weight than the negative ones regarding collocational errors (e.g., unacceptable collocations, ungrammatical collocations). However, an investigation on the effect of test takers' motivation on collocational performance (see Deci & Ryan, 2000) is beyond the scope of this study.

Table 3.10 A Summary of Operational Collocation Performance Measures (OCPMs)

Construct	OCPMs	Definition	Computation Method	Scale
Semantic accuracy	ACP_OK	# of Acceptable collocations per 500 words	500 ×# of acceptable collocations / # of words	Interval
	ACP_ERR	# of unacceptable collocations per 500 words	500 ×# of unacceptable collocations / # of words	Interval
	ACP_RAT	The ratio of acceptable to unacceptable collocations	# of acceptable collocations / (# of unacceptable collocations + 1)	Ratio
Grammatical accuracy	GRA_OK	# of error-free collocations per 500 words	500 ×# of error-free collocations / # of words	Interval
	GRA_ERR	# of erroneous collocations per 500 words	500 ×# of erroneous collocations / # of words	Interval
	GRA_RAT	The ratio of error-free collocations to erroneous collocations	# of error-free collocations / (# of erroneous collocations + 1)	Ratio
Restrictedness /precision	RES_FRE	# of highly restricted collocations per 500 words	500 ×# of highly restricted (precise) collocations / # of words	Interval
	RES_PRO	The proportion of highly restricted collocations to all collocations	# of highly restricted / # of collocations	Interval
Transparency	TRAN	The proportion of partially figurative collocations to all collocations	# of partially figurative collocations / # of collocations	Ratio
Automaticity	CHOP	The proportion of interrupted (disfluent) collocations to all collocations	# of interrupted collocations / # of collocations	Ratio
Collocational Performance Rating	CPR	A composite measure that summarizes all the collocational features of interest	20 + 2 ×# of acceptable collocations + # of substandard collocations + 2 ×# of highly restricted collocations + 2 ×# of partially figurative collocations – # of unacceptable collocations – 0.5 × # of ungrammatical collocations - 2 ×# of interrupted collocation utterances	Interval

3.4.3 Reliabilities of Collocation Coding and Oral Proficiency Measures

The results of collocation coding and oral proficiency scores (SPEAK and TEACH exam scores) investigated in this study came from subjective human observation and judgment of language behaviors based on well-conceived coding or scoring rubrics. The notion of reliability quantifies the consistency in human coding or scoring and is a fundamental consideration for behavioral measurement (Haertel, 2006). As collocation codings resulted in categorical data (see above), Fleiss' (1971) generalized Kappa was computed to examine rater agreement. The Kappa statistic discounts agreement by chance and is thus a more sophisticated measure of inter-rater agreement on nominal scales than percent agreement. Fleiss' Kappa was chosen over Cohen's Kappa (Cohen, 1960, 1968) because pairs of collocation coders were randomly selected rather than fixed (Fleiss, 1971, p. 378).

In contrast, the reliabilities of SPEAK and TEACH scores were obtained using intraclass correlation coefficients which estimate the conformity among a set of independent numerical measures on a single observation (Shrout & Fleiss, 1979). Although inter-rater reliabilities, as mentioned above, were reported by the SPEAK and TEACH testing program, they were still computed for the present dataset considering that inter-rater reliabilities may vary across testing cycles (Brown, 2012; Haertel, 2006).

3.4.4 RQ1: The Characteristics of the Collocation Occurrences in Learner Speech

To answer the first research question regarding the basic characteristics of ESL learners' collocation use in the SPEAK and TEACH exams, descriptive statistics of the collocation coding and the OCPMs were computed. In addition, a paired-sample Wilcoxon

Rank test was used to compare the medians of a few OCPMs between the SPEAK and TEACH exams, including ACP_ERR (unacceptable collocations per 500 words), GRA_ERR (erroneous collocations per 500 words), RES_FRE (highly restricted collocations per 500 words), and CHOP (the proportion of interrupted/disfluent collocations to all collocations). It was hypothesized that ESL learners had used collocations at least differently in the two distinct speaking contexts. A non-parametric paired-sample Wilcoxon Rank test was chosen because the SPEAK and TEACH responses were produced by the same group of ESL learners and the distribution of the dependent variable violates the normality assumption. All the above information gave a general picture about the sixty ESL learners' collocation usage in the two distinct communicative situations.

3.4.5 RQ2: Variation of the Collocation Measures among Proficiency Level Groups

The second research question concerned the empirical performance of each OCPM for differentiating ESL speakers of the four oral proficiency levels. Although the OCPMs were created based on theory, it was unknown whether these measures would yield different mean scores, as theory predicts, among proficiency level groups in real data. For example, it was hypothesized that the more proficient an ESL speaker was, the less frequently he or she would use unacceptable collocations in speech. Stated another way, it was expected that the mean of ACP_ERR (unacceptable collocations per 500 words) from a higher proficiency level group would be significantly smaller than that from a lower proficiency level group. To test such kind of hypotheses, a one-way between-subjects analysis of variance (ANOVA) was performed using oral proficiency level as the independent variable and each OCPM as the dependent variable. ANOVA was appropriate for this analysis of the differences in group

means because there were more than two proficiency levels in the independent variable and a dependent variable was on an interval or ratio scale.

If the OCPMs were to be adopted as scoring features in an automated scoring program, the observed variation of the mean of each OCPM among oral proficiency levels could be used as a piece of evidence for supporting the relevance of these measures to the targeted construct of academic oral proficiency, which is an important assumption underlying the evaluation inference in the hypothetical IUA mentioned in Chapter 1. That is, "Construct-relevant speech features (i.e., observed speech qualities that contribute to academic oral English proficiency) can be precisely defined" (see Chapter 1, Section 1.3.2).

3.4.6 RQ3: Prediction of Oral Proficiency based on Collocation Measurement

The third research question inquired about the relationship between the measurement outcomes of ESL learners' collocational performance in free speaking and criterion proficiency measures, including their SPEAK and TEACH exam scores and dichotomous certification decisions (certified or uncertified). To answer this question, the relationships between single collocation measures and the criterion proficiency measures were first examined using correlation analyses. Specifically, Pearson product-moment correlations were performed between OCPMs and SPEAK and TEACH scores because these variables were on an interval or ratio scale; point biserial correlations were performed between OCPMs and dichotomous certification decisions because OCPMs were on an interval or ratio scale and certification decisions were on a nominal scale. Additionally, Pearson product-moment correlations were perform between OCPMs and speech length (number of words) to provide discriminant evidence (AERA, et al., 2014, p. 17). Speech length has been found to

be a strong but not quite meaningful predictor for human ratings of oral proficiency (see e.g., Xi, et al., 2008; Xu & Xi, 2010). To discount the prospective criticism that the relationships between OCPMs and criterion proficiency measures are mediated by speech length, counterevidence to show that OCPMs are not significantly correlated with speech length must be provided.

In a sense, ANOVA and correlation yield overlapping information regarding the usefulness of OCPMs for predicting L2 oral proficiency. However, combining the two approaches is advantageous for this investigation because they address the question of prediction from different angles. ANOVA reveals group mean differences, specifically the differences in collocation usage among speakers at various oral proficiency levels. In contrast, a correlation coefficient tells the direction and overall magnitude of the association between collocation usage and oral proficiency. These two sources of information are thus complementary.

On top of the single collocation measures, a pertinent inquiry was how the single collocation measures, which surveyed every aspect of learners' oral collocational performance, as a whole, predicted the criterion proficiency measures. To conduct this inquiry, multiple regressions were performed on SPEAK and TEACH scores, respectively, as a dependent variable and OCPMs as independent variables because the dependent variable was on an interval scale. In addition, logistic regression was performed on the certification decisions as a dependent variable and OCPMs as independent variables because the dependable variable in this case was on a nominal scale. Further, step-wise procedures of selecting OCPMs were performed to find out the most parsimonious models for predicting SPEAK and TEACH scores and certification decisions.

The findings of the above inquires could be potentially used as evidence for another important assumption underlying the evaluation inference in the hypothetical IUA discussed earlier—"The way the automated features are weighted and combined to produce a holistic automated score reflects the importance of each feature to the construct of academic oral proficiency according to theory or relevant research" (see Chapter 1, Section 1.3.2). That is, if the OCPMs were to be adopted in the design of an automated scoring program for an ITA oral English assessment like the SPEAK and TEACH exams, the magnitude of associations between OCPMs and criterion proficiency measures found in this study could be used as references for evaluating the specification of the scoring model of the automated scoring program.

3.4.7 RQ4: The Context Effect on Prediction

The last research question of the study concerned the effect of context of language use on the relationship between collocational performance and perceived L2 oral proficiency. This inquiry is again relevant to the assumption regarding the accuracy of scoring model specification for predicting human criterion scores. Specifically, it investigates whether the parameters in the regression-based prediction model would vary across two different speaking contexts. To answer this question, a best-fitting regression model obtained from the SPEAK dataset were used to predict holistic oral proficiency scores in the TEACH dataset. Then, the proportion of variance in the oral proficiency scores explained by the model in the SPEAK dataset was compared with that in the TEACH dataset.

3.5 Chapter Summary

This chapter explained the details of data collection, coding, and analyses of this dissertation study. The first section discussed the source of spoken data, the sampling procedure, and the way the selected spoken data were transcribed. Then, the criteria and schemes applied to identifying and coding the collocations in the transcriptions were explicated along with the rationales. Finally, the development of meaningful operational variables of learners' collocational performance and the statistical analyses on these variables and their relationships with the criterion measures of oral proficiency were described.

CHAPTER 4: RESULTS

This chapter presents the results of data analysis performed to investigate the four research questions of this study. To answer the first research question on the basic characteristics of the collocations in learner speech, reliability estimates of collocation coding and descriptive statistics of the collocation measures were computed. To answer the second research question about the differences in the collocation measures among proficiency level groups, analysis of variance (ANOVA) was performed. To answer the third research question regarding how well the collocation measures predict human criterion measures of oral proficiency, correlation and regression analyses were conducted. To answer the last question or test the hypothesis that the relationship between collocation and L2 oral proficiency interacts with speaking contexts, a cross-validation technique was adopted.

4.1 The Characteristics of the Collocation Occurrences in Learner Speech

The first research question investigated the basic characteristics of the collocation occurrences in the sixty ESL learners' SPEAK and TEACH responses. To answer this question, reliability and descriptive statistics analysis of the collocation measures was conducted. In addition, paired-sample Wilcoxon Rank tests were performed to examine the differences in a few measures between the two exams.

4.1.1 Reliability Estimates of Collocation Extraction and Coding

In this section, the reliability estimates of manual collocation extraction, criterion measures of L2 oral proficiency, and paired collocation coding on semantic accuracy, grammatical accuracy, restrictedness, and transparency are reported.

4.1.1.1 Precision and Recall of Collocation Extraction

Precision and recall of a single coder's collocation extraction were calculated based on a pilot sample of twenty speakers, which was a subset of the complete sample, and by comparing the researcher's extraction with the true collocations agreed upon by both the researcher and a second coder (see formulas in Chapter 3, Section 3.3.1). The precision rate based on 391 true collocations (approximately 17% of the total collocations identified in this study) in the pilot sample was .99, indicating that 99% of the collocations that the researcher identified alone were true collocations. The recall rate was .95, indicating that he only missed 5% of the true collocations in the transcriptions. These statistics suggest that collocation extraction performed by a single coder is sufficiently accurate.

4.1.1.2 Inter-coder and Inter-rater Reliabilities

Reliability estimates were calculated for both collocation coding (from which the collocation measures of this study were derived) and the criterion measures of oral proficiency (i.e., SPEAK and TEACH exam ratings). Table 4.1 displays the levels of inter-coder agreement achieved for the five collocation coding subcategories. The percent agreement (percentage of absolute agreement between two coders) ranged from 73.7% to 93.0%; Kappa statistics ranged from .262 to .657, indicating fair to substantial agreement (Landis & Koch, 1977). These statistics, however, were relatively low compared to inter-coder Kappa statistics reported by other language testing studies. For example, in Brown, et al.'s (2005) study, the inter-coder agreement on coding six conceptual categories of human raters' self-reported rating orientations was between .79 to .94; in Xi et al.'s (2008) study, the human agreement on rating spoken responses in four categories was between .54 to .71. The

lowest inter-coder agreement in this study was on semantic accuracy and restrictedness. Possible causes of the low inter-coder agreement will be discussed in the next chapter. The reliability estimates of the SPEAK and TEACH ratings were computed using intraclass correlation coefficients of three concurrent raters. The ICCs for SPEAK and TEACH were .918 and .924, respectively, indicating outstanding agreement among the human raters.

Table 4.1 Inter-coder Agreement by Collocation Coding Categories (n = 60)

	SPEAK		**TEACH**	
	% agreement	**Kappa**	**% agreement**	**Kappa**
Semantic accuracy	73.7%	.379	76.1%	.254
Grammatical accuracy	88.7%	.657	91.2%	.568
Transparency	85.6%	.398	93.0%	.533
Restrictedness	76.9%	.262	74.2%	.470
Automaticity	N/A	N/A	N/A	N/A

Note: Automaticity was coded by a single coder

4.1.2 Descriptive Statistics

From the sixty participants' transcribed spoken responses in the SPEAK and TEACH exams, a total of 2344 collocation strings (1272 from SPEAK and 1072 from TEACH) were extracted and coded. Based on the final codings (after adjudication), a majority of these collocations were acceptable (62.7% in SPEAK and 73.8% in TEACH), error-free (75.0% in SPEAK and 92.3% in TEACH), literal (92.5% in SPEAK and 93.6% in TEACH), moderately restricted (80.3% in SPEAK and 57.5% in TEACH), and uttered coherently (75.9% in SPEAK and 79.8% in TEACH); in general, the speakers' collocational performance appeared to be better in the TEACH exam than in the SPEAK exam—there were overall more acceptable, error-free, and highly restricted collocations in the TEACH responses than in the SPEAK responses (see Table 4.2). Additionally, the speakers used

verb-noun collocations the most frequently, adjective-noun collocations the second, and adverb-adjective, phrasal verb-adverb, and noun-phrasal verb patterns the least frequently; in general, semantic accuracy rate (i.e., the proportion of acceptable collocations) seemed to be higher in TEACH than in SPEAK in almost all syntactic patterns except phrasal verb-noun (see Table 4.3).

Table 4.2 Descriptive Statistics of the Coded Collocations by Coding Categories (n = 60)

Variable	SPEAK	TEACH
Acceptable	797 (62.7%)	791 (73.8%)
Substandard	256 (20.1%)	125 (11.7%)
Unacceptable	219 (17.2%)	156 (14.5%)
Error-free	954 (75.0%)	989 (92.3%)
Erroneous	318 (25.0%)	83 (7.7%)
Partially figurative	96 (7.5%)	69 (6.4%)
Literal	1176 (92.5%)	1003 (93.6%)
Highly restricted	250 (19.7%)	456 (42.5%)
Moderately restricted	1022 (80.3%)	616 (57.5%)
Uninterrupted	965 (75.9%)	855 (79.8%)
Interrupted	307 (24.1%)	217 (20.2%)
Total	1272	1072

Note: % in the parenthesis stands for the number divided by the total number of collocations.

Table 4.3 Descriptive Statistics of Frequency and Semantic Accuracy of Extracted Collocations by Syntactic Pattern (n = 60)

Pattern	SPEAK		TEACH		TOTAL	
	Count	Acceptable (%)	Count	Acceptable (%)	Count	Acceptable (%)
ADJ-N	305	198 (64.9%)	282	223 (79.1%)	587	421 (71.7%)
ADV-ADJ	4	4	6	4	10	8
ADV-V	47	37 (78.7%)	40	33 (82.5%)	87	70 (80.5%)
N-N	135	87 (64.4%)	110	82 (74.5%)	245	169 (69.0%)
N-of-N	43	24 (55.8%)	110	84 (76.4%)	153	108 (70.6%)
N-V	12	8	32	27	44	35
V-N	601	361 (60.1%)	431	308 (71.5%)	1032	669 (64.8%)
PHV-ADV	8	6	1	1	9	7
N-PHV	8	5	4	2	12	7
PHV-N	109	67 (61.5%)	56	27 (48.2%)	165	94 (57.0%)
Total	1272	797 (62.7%)	1072	791 (73.8%)	2344	1588 (67.7%)

Note: % stands for the accurate rate for each syntactic pattern

The descriptive statistics of the operational collocational performance measures (OCPMs), as presented in Table 4.4, show the profile of an average Chinese ITA's collocational performance in naturally occurring speech. In the SPEAK exam, a speaker, on average, produced approximately twenty collocations per 500 words; among them, thirteen were acceptable collocations, fifteen were surface-error-free collocations, and four were highly restricted (precise) collocations; at the same time, the speaker made a small number of collocational errors: approximately four unacceptable collocations and five ungrammatical collocations; the utterance of the collocations were in many cases (three thirds) uninterrupted. In the TEACH responses, the numbers were quite similar except that a speaker, on average, produced two fewer unacceptable collocations, four fewer surface-error-free collocations, and three more highly restricted collocations.

Table 4.4 Descriptive Statistics of Collocation Performance Measures (n=60)

Variables	**SPEAK**			**TEACH**		
	Mean	***SD***	***Range***	***Mean***	***SD***	***Range***
Average speech length (# of words)	523.40	122.16	215-756	528.63	101.40	253-733
Total # per 500 words	20.42	5.11	10.00-31.67	16.70	7.77	3.80-34.38
ACP_OK: # of acceptable collocations per 500 words	12.66	5.02	3.96-26.10	12.42	6.23	1.66-30.45
ACP_ERR: # of unacceptable collocations per 500 words	3.61	2.82	.00-11.70	2.34	2.87	.00-15.93
ACP_RAT: Proportion of acceptable to unacceptable collocations	5.57	6.22	.56-25.00	6.86	6.22	.67-26.00
GRA_OK: # of error-free collocations per 500 words	15.39	5.60	5.35-27.83	15.38	7.17	2.53-33.71
GRA_ERR: # of erroneous collocations per 500 words	5.03	2.71	.89-11.01	1.32	1.58	.00-8.49

Table 4.4 (continued)

Variables	SPEAK			TEACH		
	Mean	***SD***	***Range***	***Mean***	***SD***	***Range***
GRA_RAT: Proportion of error-free to erroneous collocations	3.60	3.19	.64-14.00	8.69	5.96	1.00-36.00
RES_FRE: # of highly restricted collocations per 500 words	4.10	2.71	.00-12.90	7.14	.89	.00-21.61
RES_PRO: Proportion of highly restricted to all collocations	.20	.11	.00-.45	.42	.18	.00-.85
TRAN: Proportion of partially figurative to all collocations	.08	.07	.00-.28	.06	.07	.00-.29
CHOP: Proportion of disfluent to all collocations	.25	.20	.00-.78	.25	.23	.00-.85
CPR: Collocational performance rating	44.98	21.89	5.00-97.50	54.55	25.27	15.00-126.00

4.1.3 Wilcox Rank Test

To find out whether there was a statistical difference in the Chinese ITAs' use of unacceptable collocations, erroneous collocations, and highly restricted collocations in the SPEAK and TEACH exams, paired-sample Wilcoxon Rank tests were performed using exam type (SPEAK or TEACH) as an independent variable and ACP_EER (the number of unacceptable collocations per 500 words), GRA_ERR (the number of erroneous collocations per 500 words), and RES_FRE (the number of highly restricted collocations per 500 words), respectively, as a dependent variable.

A Wilcoxon Rank test was chosen over a T-test because the distributions of the above dependent variables violated the normality assumption (Table 4.5). A paired-sample Wilcoxon Rank test is a non-parametric equivalent to a paired-sample T-test. Different from a T-test that compares the means of two variables, a Wilcoxon Rank test compares the

medians of two variables. Being non-parametric means that normality of a dependent variable is not assumed in this analysis.

The results of the Wilcoxon Rank tests suggested that the learners tended to use significantly fewer unacceptable collocations (Z = -2.968, p < .01) and ungrammatical collocations (Z = -6.331, p < .01) and more highly restricted collocations (Z = -4.203, p < .01) in TEACH exam than in the SPEAK exam.

Table 4.5 Tests of Normality of the ACP_OK, ACP_ERR, and ACP_RAT in the SPEAK and TEACH Datasets (n=60)

Variable	Exam	Shapiro-Wilk *W*	*df*	*p*
ACP_ERR	SPEAK	.935	60	.003
	TEACH	.772	60	.000
GRA_ERR	SPEAK	.963	60	.065
	TEACH	.762	60	.000
RES_FRE	SPEAK	.946	60	.010
	TEACH	.934	60	.003

4.2 Variation of the Collocation Measures among Proficiency Level Groups

One primary goal of this study was to identify OCPMs that were useful for differentiating among Chinese ITAs of different oral proficiency levels. An ideal collocational measure should yield significantly different means across the four oral proficiency levels; in addition, the differences in the measurement outcomes should be interpretable based on theory. To answer the second research question of whether collocation measures differ across proficiency level groups, a one-way between-subjects analysis of variance (ANOVA) was performed on each OCPM as a function of the speakers' oral English proficiency levels. As the speakers' proficiency levels were determined based on

both SPEAK and TEACH exams, OCPMs were obtained from the aggregated SPEAK and TEACH responses.

4.2.1 Measures on Semantic Accuracy

The three measures derived from the human coding on semantic accuracy (i.e., the meaningfulness of the combination of two content words) were ACP_OK (normalized frequency of acceptable collocations), ACP_ERR (normalized frequency of unacceptable collocations), and ACP_RAT (proportion of acceptable to unacceptable collocations). It is generally assumed that acceptable collocations enhance speech comprehensibility whereas unacceptable collocations obscure the meaning of speech. It was thus expected that more proficient speakers would use acceptable collocations more frequently and unacceptable collocations less frequently. Thus, ACP_OK and ACP_RAT were expected to yield larger group means in higher proficiency groups whereas ACP_ERR to yield smaller group means in higher proficiency groups.

As the first step of the analysis, three important assumptions of ANOVA (normality of errors/residuals, homogeneity of variance, outliers/influential cases) were checked. The assumption of normality was met for all levels except Level 1 in ACP_OK and for all levels except Level 4 in ACP_ERR; due to the violation of the normality assumption in most levels in ACP_RAT, the variable was inverted following Tabachnick and Fidell's (2005) data transformation guidelines. The transformed variable InvACP_RAT was equally interpretable—the ratio of unacceptable to acceptable collocations (Table 4.6). The assumption of homogeneity of variance was met in ACP_OK (Brown-Forsythe $F(3, 56)$

=.347, p = .792) but violated in ACP_ERR (Brown-Forsythe $F(3, 56)$ =3.496, $p < .05$) and ACP_RAT (Brown-Forsythe $F(3, 56)$ =4.555, $p < .01$). No obvious outliers were spotted.

Table 4.6 Tests of Normality of the ACP_OK, ACP_ERR, and ACP_RAT Scores in Each Oral Proficiency Level (n=60)

Variable	Level	Shapiro-Wilk *W*	*df*	*p*
ACP_OK	1 (Fully certified)	.866	15	.030
	2 (Borderline)	.922	15	.209
	3 (Limited)	.941	15	.389
	4 (Very Limited)	.944	15	.442
ACP_ERR	1 (Fully certified)	.962	15	.732
	2 (Borderline)	.891	15	.069
	3 (Limited)	.893	15	.076
	4 (Very Limited)	.862	15	.026
Inv(ACP_RAT)	1 (Fully certified)	.911	15	.138
	2 (Borderline)	.786	15	.002
	3 (Limited)	.915	15	.161
	4 (Very Limited)	.958	15	.651

A significant effect of oral proficiency levels was found for ACP_OK ($F(3, 56)$ = 10.276, $p < .01$, η^2 = .355), ACP_ERR ($F(3, 56)$ = 12.707, $p < .01$, η^2 = .405), and InvACP_RAT ($F(3, 56)$ = 26.792., $p < .01$, η^2 = .589). Post hoc analysis on ACP_OK using the Scheffé adjustment indicated that highly proficient speakers on average used acceptable collocations more frequently than less proficient speakers (Level 1 vs. Level 4, $p < .01$; Level 2 vs. Level 3, $p < .05$; Level 2 vs. Level 4, $p < .01$; see Table 4.6). Additionally, post hoc pairwise comparisons were performed on ACP_ERR and Inverse of ACP_RAT, respectively, using the Games-Howell procedure for unequal variances across levels. Results indicated that high-level speakers on average produced unacceptable collocations less frequently than low-level speakers (Level 2 vs. Level 3, $p < .01$; Level 2 vs. Level 4, $p < .01$); further, the former

had lower deviant-to-sound collocation ratio in speech than the latter (Level 1 vs. Level 3, $p < .01$; Level 1 vs. Level 4, $p < .01$; Level 2 vs. Level 3, $p < .01$; Level 2 vs. Level 4, $p < .01$; see Table 4.7).

Table 4.7 Descriptive Statistics of ACP_OK, ACP_ERR, and Inv(ACP_RAT) Scores across Oral Proficiency Levels (n=60)

Variable	Level	*N*	*Mean*	*SD*
ACP_OK	1 (Fully certified)	15	14.63	4.01
	2 (Borderline)	15	15.81	4.72
	3 (Limited)	15	10.98	3.90
	4 (Very Limited)	15	8.73	2.97
ACP_ERR	1 (Fully certified)	15	2.92	1.78
	2 (Borderline)	15	1.74	1.67
	3 (Limited)	15	4.79	2.46
	4 (Very Limited)	15	3.47	1.94
Inv(ACP_RAT)	1 (Fully certified)	15	.17	.08
	2 (Borderline)	15	.18	.04
	3 (Limited)	15	.33	.13
	4 (Very Limited)	15	.28	.12

4.2.2 Measures on Grammatical Accuracy

Similarly, it was hypothesized that the more proficient a speaker was, the fewer erroneous collocations he or she would produce in spontaneous speech. The OCPMs related to grammatical accuracy were GRA_OK (normalized frequency of error-free collocations), GRA_ERR (normalized frequency of erroneous collocations), and GRA_RAT (the ratio of error-free to erroneous collocations). The same ANOVA procedure was performed on these variables.

First, the assumptions of ANOVA were checked. The assumption of normality was satisfied for all levels in GRA_OK, for all levels except Level 4 in GRA_ERR, and for all levels except Level 3 in GRA_RAT. The assumption of homogeneity of variance was met in GRA_OK (Brown-Forsythe $F(3, 56) = 1.643$, $p = .190$) and GRA_ERR (Brown-Forsythe $F(3, 56) = 1.371$, $p = .261$) but violated in GRA_RAT (Brown-Forsythe $F(3, 56) = 3.715$, $p < .01$). No obvious outliers were spotted (see Table 4.8). When the homogeneity assumption is not met, there is a chance that the null hypothesis is being falsely rejected. However, it has been found that ANOVA is fairly robust to the violation to this assumption (Tomarken & Serlin, 1986).

Table 4.8 Tests of Normality of the GRA_OK, GRA_ERR, and GRA_RAT Scores in Each Proficiency Level (n=60)

Variable	Level	Shapiro-Wilk *W*	*df*	*p*
GRA_OK	1 (Fully certified)	.921	15	.202
	2 (Borderline)	.951	15	.540
	3 (Limited)	.929	15	.265
	4 (Very Limited)	.931	15	.284
GRA_ERR	1 (Fully certified)	.957	15	.633
	2 (Borderline)	.908	15	.128
	3 (Limited)	.936	15	.335
	4 (Very Limited)	.866	15	.029
GRA_RAT	1 (Fully certified)	.916	15	.167
	2 (Borderline)	.922	15	.210
	3 (Limited)	.832	15	.010
	4 (Very Limited)	.923	15	.217

As expected, oral proficiency levels was found to have a significant effect on GRA_OK ($F(3, 56) = 6.484$, $p < .01$, $\eta^2 = .258$), GRA_ERR ($F(3, 56) = 15.791$, $p < .01$, $\eta^2 = .458$), and GRA_RAT ($F(3, 56) = 7.827.$, $p < .01$, $\eta^2 = .295$), respectively. Post hoc analysis

on GRA_OK using the Scheffé adjustment indicated that highly proficient speaker appeared to use more grammatically accurate collocations than less proficient speakers in speech (Level 1 vs. Level 3, $p < .01$; Level 1 vs. Level 4, $p < .01$). The same analysis performed on GRA_ERR suggested that the most and the least proficient groups tended to use significantly fewer grammatically erroneous collocations than the two groups in the middle (Level 1 vs. Level 2, $p < .01$; Level 1 vs. Level 3; $p < 01$; Level 2 vs. Level 4, $p < .01$; Level 3 vs. Level 4, $p < .01$; see Table 4.8 for means). Post hoc analysis on GRA_RAT, using the Games-Howell procedure for unequal variances across levels, revealed that the highest level proficiency group had significantly higher ratio of error-free to erroneous collocations in speech than the other three groups (Level 1 vs. Level 2, $p < .01$; Level 1 vs. Level 3, $p < .01$; Level 1 vs. Level 4, $p < .01$; see Table 4.9).

Table 4.9 Descriptive Statistics of GRA_OK, GRA_ERR, and GRA_RAT Scores across Proficiency Level Groups (n=60)

Variable	**Level**	***n***	***Mean***	***SD***
GRA_OK	1 (Fully certified)	15	19.17	5.40
	2 (Borderline)	15	16.32	4.69
	3 (Limited)	15	13.42	4.69
	4 (Very Limited)	15	12.62	3.04
GRA_ERR	1 (Fully certified)	15	2.01	.94
	2 (Borderline)	15	4.03	1.63
	3 (Limited)	15	4.56	1.24
	4 (Very Limited)	15	2.08	1.22
GRA_RAT	1 (Fully certified)	15	9.55	4.60
	2 (Borderline)	15	5.23	2.79
	3 (Limited)	15	4.97	2.70
	4 (Very Limited)	15	4.93	1.86

4.2.3 Measures on Restrictedness or Precision

The overall restrictedness or precision of collocation production was measured by RES_FRE (the normalized frequency of highly restricted collocations) and RES_PRO (the proportion of highly restricted collocations to all collocations produced). It was anticipated that high-level speakers would produce more highly restricted (precise) collocations than low-level speakers.

The ANOVA assumptions were checked before conducting analyses. The assumption of normality was met for all levels in RES_FRE and RES_PRO (see Table 4.10). The assumption of homogeneity of variance also met in RES_FRE (Brown-Forsythe $F(3, 56) = 1.489$, $p = .227$) and RES_PRO (Brown-Forsythe $F(3, 56) = .298$, $p = .827$). No outliers were observed.

Table 4.10 Tests of Normality of RES_FRE and RES_PRO Scores in Each Oral Proficiency Level (n=60)

Variable	Level	Shapiro-Wilk *W*	*df*	*p*
RES_FRE	1 (Fully certified)	.962	15	.730
	2 (Borderline)	.929	15	.261
	3 (Limited)	.946	15	.470
	4 (Very Limited)	.926	15	.236
RES_PRO	1 (Fully certified)	.976	15	.932
	2 (Borderline)	.938	15	.361
	3 (Limited)	.941	15	.396
	4 (Very Limited)	.976	15	.936

ANOVA results suggested a significant difference in RES_FRE ($F(3, 56) = 6.890$, $p < .01$, $\eta^2 = .270$) and RES_PRO ($F(3, 56) = 4.063$, $p < .05$, $\eta^2 = .179$) across oral proficiency levels. Post hoc analyses on RES_FRE and RES_PRO using the Scheffé adjustment

indicated that Level 2 speakers used highly restricted collocations more frequently than Level 3 ($p < .01$) as well as Level 4 speakers ($p < .05$); in addition, a larger proportion of highly restricted collocations was found in Level 2 speakers' collocation production than in Level 3 speakers' collocation production ($p < .05$). No other significant differences between group means were found.

Table 4.11 Descriptive Statistics of RES_FRE and RES_PRO Scores across Proficiency Level Groups (n=60)

Variable	Level	*n*	*Mean*	*SD*
RES_FRE	1 (Fully certified)	15	6.44	2.98
	2 (Borderline)	15	7.59	3.10
	3 (Limited)	15	4.04	1.89
	4 (Very Limited)	15	4.41	1.61
RES_PRO	1 (Fully certified)	15	.31	.09
	2 (Borderline)	15	.37	.08
	3 (Limited)	15	.25	.10
	4 (Very Limited)	15	.32	.10

4.2.4 Measures on Transparency

The OCPM regarding the transparency of a speaker's collocation production was TRAN, i.e., the proportion of partially figurative collocations to all collocational occurrences in speech. It was expected that advanced speakers would use a larger proportion of partially figurative collocations in speaking as figurative language use assumingly increases the native-likeness of the speech.

The assumptions of ANOVA were first checked. The assumption of normality was met for all levels except Level 3 (see Table 4.12). The assumption of homogeneity of variance was, however, not met (Brown-Forsythe $F(3, 56) = 4.237$, $p < .01$). No obvious

outliers were spotted. Based on the ANOVA results, there was no significant difference in the proportion of partially figurative across oral proficiency levels, $F(3, 56) = 2.704$, $p = .054$, $\eta^2 = .102$.

Table 4.12 Tests of Normality of TRAN Scores in Each Proficiency Level (n=60)

Variable	Level	Shapiro-Wilk *W*	*df*	*p*
TRAN	1 (Fully certified)	.908	15	.908
	2 (Borderline)	.928	15	.258
	3 (Limited)	.809	15	.005
	4 (Very Limited)	.899	15	.051

Table 4.13 Descriptive Statistics of TRAN Scores across Proficiency Level Groups (n=60)

	Level	*n*	*Mean*	*SD*
TRAN	1 (Fully certified)	15	.08	.04
	2 (Borderline)	15	.07	.06
	3 (Limited)	15	.06	.04
	4 (Very Limited)	15	.06	.05

4.2.5 Measures on Automaticity

Automaticity was measured by CHOP or the proportion of interrupted (disfluent) collocations among all collocational occurrences in speech. As discussed in Chapter 2, Section 2.3.5, collocations are believed to promote speech fluency. Thus, it was anticipated that advanced speakers would produce fewer chopped collocations than low-level speakers. To test this hypothesis, a one-way between-subjects ANOVA was conducted on CHOP as a function of the oral proficiency levels.

The assumptions of ANOVA were examined before the analysis. The assumption of normality was met for all levels except Level 2 (Table 4.14). The assumption of homogeneity

of variance was, however, not met (Brown-Forsythe $F(3, 56) = 5.733$, $p < .05$). No outliers were detected.

Table 4.14 Tests of Normality of CHOP Scores in Each Proficiency Level (n=60)

Variable	Level	Shapiro-Wilk *W*	*df*	*p*
CHOP	1 (Fully certified)	.928	15	.251
	2 (Borderline)	.871	15	.035
	3 (Limited)	.943	15	.416
	4 (Very Limited)	.956	15	.626

The ANOVA results suggested a significant difference in the proportion of interrupted (disfluent) collocations across oral proficiency levels, $F(3, 56) = 41.230$, $p < .01$, $\eta^2 = .688$. Post hoc analysis, using the Games-Howell procedure for unequal variances across levels revealed that Level 2 speakers had the least proportion of disfluent collocations, followed by Level 1 ($p < .01$), Level 3 ($p < .01$), and Level 4 ($p < .01$). This indicated that Level 2 speakers were most fluent in uttering collocations whereas Level 4 speakers were the least fluent.

Table 4.15 Descriptive Statistics of CHOP Scores across Proficiency Level Groups (n=60)

Variable	Level	*n*	*Mean*	*SD*
CHOP	1 (Fully certified)	15	.15	.05
	2 (Borderline)	15	.07	.07
	3 (Limited)	15	.29	.12
	4 (Very Limited)	15	.50	.17

4.2.6 The Composite Measure

Finally, the empirical performance of the composite measure, namely Collocational Performance Rating or CPR, was examined. As the measure adds up positive collocational

features and deducts negative ones (see Chapter 3, Section 3.3.4), it was anticipated that more proficient speakers would obtain higher CPR scores. A one-way between-subjects ANOVA was conducted on CPR as a function of oral proficiency levels.

The assumptions of ANOVA were checked. The assumption of normality was met for all proficiency levels (Table 4.16). The assumption of homogeneity of variance was also met (Brown-Forsythe $F(3, 56) = 2.372$, $p = .080$). No outliers were spotted.

There was a significant difference in CPR across oral proficiency levels, $F(3, 56) = 44.594$, $p < .01$, $\eta^2 = .705$. Post hoc Scheffé analysis suggested that high-level speakers yielded a significantly higher CPR value than low-level speakers (Level 1 vs. Level 3, $p < .01$; Level 1 vs. Level 4, $p < .01$; Level 2 vs. Level 3, $p < .01$; Level 3 vs. Level 4, $p < .01$; no other significant differences between levels were found). The descriptive statistics of CPR across proficiency levels can be found in Table 4.17.

Table 4.16 Tests of Normality of CPR Scores in Proficiency Level Groups (n=60)

Variable	Level	Shapiro-Wilk *W*	*df*	*p*
CPR	1 (Fully certified)	.927	15	.246
	2 (Borderline)	.960	15	.690
	3 (Limited)	.953	15	.577
	4 (Very Limited)	.961	15	.710

Table 4.17 Descriptive Statistics of CPR Scores across Proficiency Level Groups (n=60)

Variable	Level	*n*	*Mean*	*SD*
CPR	1 (Fully certified)	15	64.26	14.96
	2 (Borderline)	15	70.08	14.66
	3 (Limited)	15	35.58	8.94
	4 (Very Limited)	15	29.13	6.49

4.2.7 A Summary of Promising Collocation Measures

In sum, a majority of OCPMs were found to vary significantly across oral proficiency levels. They included ACP_OK (the number of acceptable collocations per 500 words), ACP_ERR (the number of unacceptable collocations per 500 words), Inv (ACP_RAT) (the ratio of unacceptable to acceptable collocations), GRA_OK (the number of error-free collocations per 500 words), GRA_ERR (the number of erroneous collocations per 500 words), GRA_RAT (the ratio of error-free to erroneous collocations), RES_FRE (the number of highly restricted collocations per 500 words), RES_PRO (the proportion of highly restricted collocations to all collocations), CHOP (the proportion of interrupted (disfluent) collocations to all collocations), and CPR (the composite measure).

4.3 Prediction of Oral Proficiency based on Collocation Measures

The third research question investigated how well the collocation measures predicted human criterion scores of oral proficiency and the dichotomous certification decisions. To answer this question, correlational analysis, multiple and logistic regressions were performed. Further, the predictive power between the composite scoring approach (CPR) and the regression approach were compared.

4.3.1 Correlational Analyses

Correlational analyses revealed how each OCPM alone predicted the human criterion scores of oral English proficiency. The collocation measures were interval data and the oral proficiency measures included both interval data (test scores resulting from human ratings) and ordinal data (certification decisions). For this reason, both Pearson product-moment

correlation and point biserial correlation coefficients were computed. The point biserial correlation coefficient or r_{pb} is used to measure the association or relationship between a continuous variable, in this case the SPEAK or TEACH scores, and a dichotomous variable, in this case, the certification decisions. The four oral proficiency levels were dichotomized into two levels: certified or uncertified. Level 1 (fully certified) and Level 2 (borderline) were coded as the certified group while Level 3 (limited) and Level 4 (very limited) were coded as the uncertified group (see Chapter 3, Section 3.1.1).

The correlational analyses were preceded by an examination of the assumptions, including linearity, range of the data (a restricted range of the score would attenuate the magnitude of the correlation coefficient), extreme data points, and the reliability of the measures (which have been reported above). The scatter plots shown in Figure 4.1 and Figure 4.2 indicate that the bivariate relationships between OCPMs and SPEAK or TEACH scores were generally linear and that the range of the data on either measure was sufficiently large; in addition, no extreme values were spotted.

Table 4.18 presents the Pearson and point biserial correlation coefficients between the collocation measures and the human criterion scores of oral proficiency. It was found that a majority of the collocation measures by themselves were significant, either moderate or strong predictors for the human criterion scores of oral proficiency. The strong predictors were CPR (the composite collocational performance measure) and CHOP (proportion of disfluent collocations). The moderate predictors were ACP_OK (normalized frequency of acceptable collocations), ACP_ERR (normalized frequency of unacceptable collocations), ACP_RAT (ratio of acceptable to unacceptable collocations), GRA_OK (normalized frequency of error-free collocations), GRA_RAT (ratio of error-free to erroneous

collocations), and RES_FRE (normalized frequency of highly restricted collocations). The remaining measures, including GRA_ERR (normalized frequency of erroneous collocations), RES_PRO (proportion of highly restricted collocations), and TRAN (proportion of partially figurative collocations) were not found to be significant predictors.

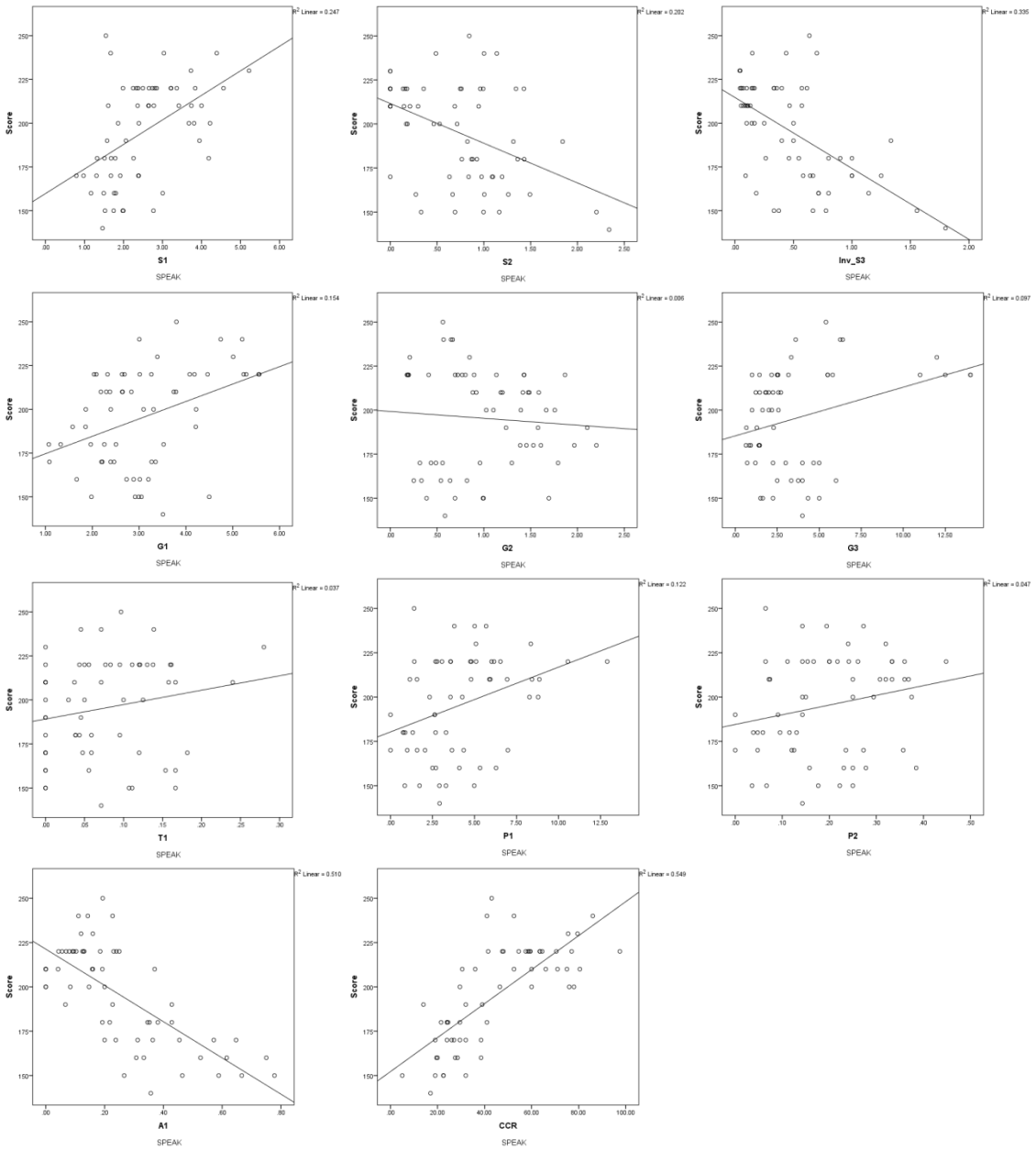


Figure 4.1. The scatterplots of OCPMs and SPEAK scores

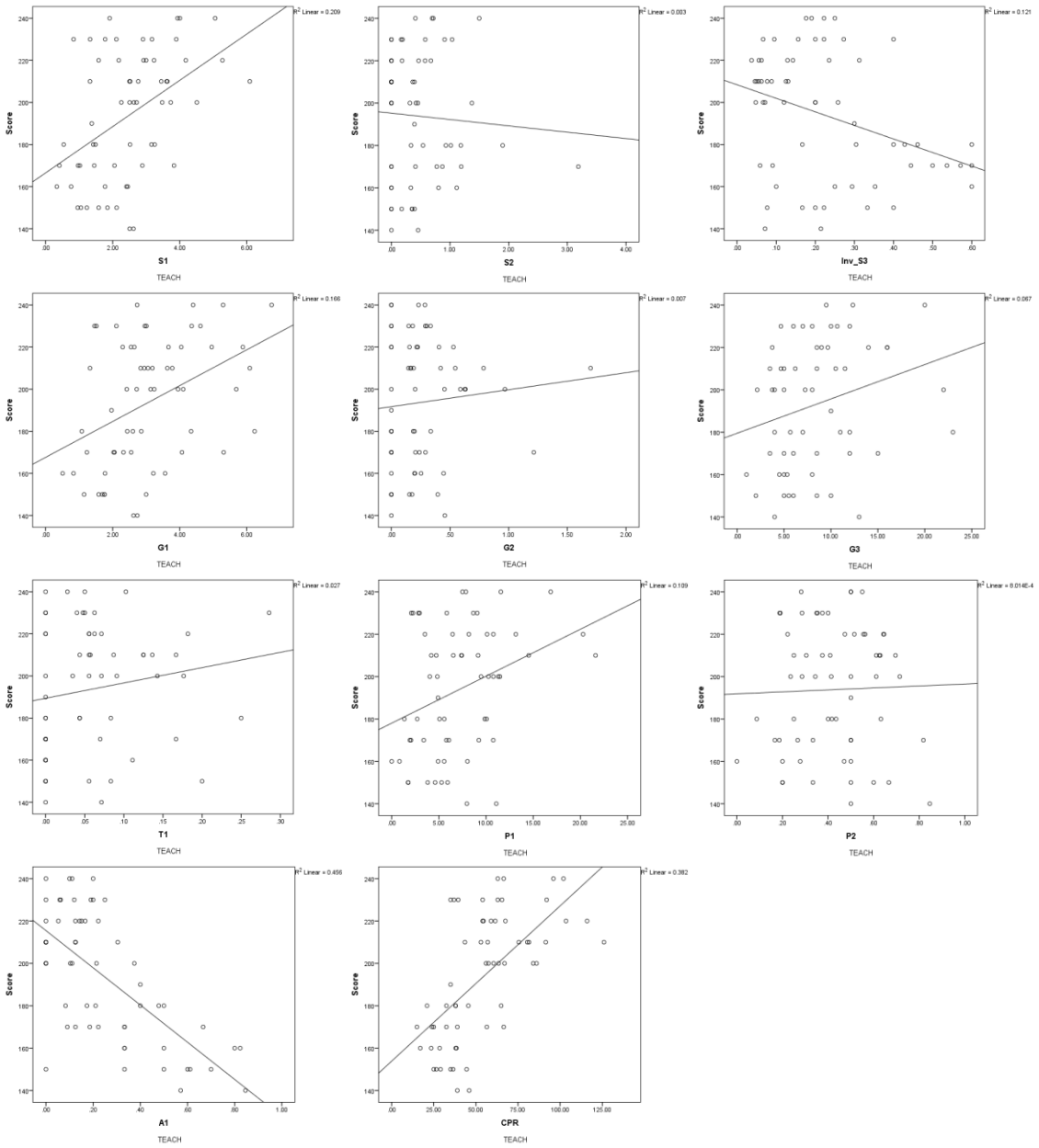


Figure 4.2. The scatterplots of OCPMs and TEACH scores

In addition, the correlations between the collocation measures and the length of speech were examined to discount the hypothesis that the relationships between collocation measures and oral proficiency measures could be mediated by speech length. The results

suggested that none of the collocation measure had a significant strong relationship with speech length.

As above, the single collocation measures seemed to predict both certification decisions and human criterion scores oral English proficiency well. The next section will investigate to what extent these measures, when combined, would predict a speaker's oral English proficiency.

Table 4.18 Pearson and Correlations between OCPMs and SPEAK/TEACH Scores, Placement and Speech Lengths (n=60)

OCPMs	SPEAK			TEACH		
	Rater Score	**Certified or not**	**Length**	**Rater Score**	**Certified or not**	**Length**
ACP_OK: # of acceptable collocations per 100 words	.497**	.517**	-.056	.457**	.398**	.140
ACP_ERR: # of unacceptable collocations per 100 words	-.449**	-.531**	-.146	-.057	-.197	.013
ACP_RAT: Ratio of acceptable to unacceptable collocations	.463**	.561**	.140	.248	.337**	.177
GRA_OK: # of error-free collocations per 100 words	.393**	.446**	-.130	.407**	.286*	.146
GRA_ERR: # of erroneous collocations per 100 words	-.075	-.195	.002	.085	.060	.017
GRA_RAT: Ratio of error-free to erroneous collocations	.311**	.353**	.002	.317*	.221	.242
RES_FRE: # of highly restricted collocations per 100 words	.349**	.539**	-.097	.331**	.304*	.043
RES_PRO: Proportion of highly restricted to all collocations	.216	.446**	-.071	.028	.091	-.033
TRAN: Proportion of partially figurative to all collocations	.193	.283	.162	.164	.157	-.031
CHOP: Proportion of disfluent to all collocations	-.714**	-.687**	-.224	-.675**	-.526**	-.145
CPR: Collocational performance rating	.741**	.825**	.261**	.618**	N/A	.353**

Note: ** $p < .01$; * $p < .05$

4.3.2 Multiple Regression for Predicting SPEAK Score

Regression analysis was first performed on the SPEAK exam dataset. A standard multiple regression was performed on final SPEAK score predicted by ACP_OK (normalized frequency of acceptable collocations), ACP_ERR (normalized frequency of unacceptable collocations), ACP_RAT (the ratio of acceptable to unacceptable collocations), GRA_OK (normalized frequency of error-free collocations), GRA_ERR (normalized frequency of erroneous collocations), GRA_RAT (the ratio of error-free to erroneous collocations), TRAN (the proportion of partially figurative collocations), RES_FRE (normalized frequency of highly restricted collocations), RES_PRO (the proportion of highly restricted collocations), and CHOP (the proportion of disfluent collocations).

Before conducting the analysis, the assumptions of a standard multiple regression were checked. Collinearity statistics indicated that multicollinearity was a concern to the analysis. Specifically, the variance inflation factor (VIF) values of GRA_OK, RES_FRE, and RES_PRO, were over 10, the rule-of-thumb cut-off value (Table 4.19). Considering that GRA_OK, GRA_ERR, and GRA_RAT were created to measure the same construct of grammatical accuracy and RES_FRE and RES_PRO were created to measure the same construct of restrictedness/precision, GRA_OK and RES_PRO were removed from the regression model to reduce predictor repetitiveness. In a regression model, if two predictor variables measure the same construct, they are likely to be highly correlated (Table 4.20). Usually, keeping both of them in the model is redundant, meaning that the additional predictor variable would not significantly increase the effect size of the model. So a common way to deal with multicollinearity is removing the repetitive predictor variables.

All the VIFs in the reduced model were less than 10, suggesting that multicollinearity was no longer a concern. Three outliers (standardized residuals exceeding |2|) were deleted from analysis. The assumption of normality and homoscedasticity were met: Shapiro-Wilk $W = .966$, $p = .094$, Breusch-Pagan $\chi^2(8) = 3.902$, $p = .866$, respectively. Linearity also seemed to hold because the scatter cloud as shown in the scatterplot of standardized residuals (Figure 4.3) is around the horizontal line of zero. All other assumptions were met. The descriptive and correlation matrix among the variables are displayed in Table 4.20.

Table 4.19 Collinearity Statistics for the SPEAK Dataset (n=60)

Variable	**VIF**	
	Complete Model	**Model with GRA_OK and RES_PRO removed**
ACP_OK	7.284	3.118
ACP_ERR	4.184	2.688
ACP_RAT	3.421	3.110
GRA_OK	10.743*	N/A
GRA_ERR	6.063	3.955
GRA_RAT	4.910	4.185
TRAN	1.214	1.147
RES_FRE	26.323*	1.465
RES_PRO	19.581*	N/A
CHOP	1.690	2.110

Note: * VIF > the cut-off value of 10

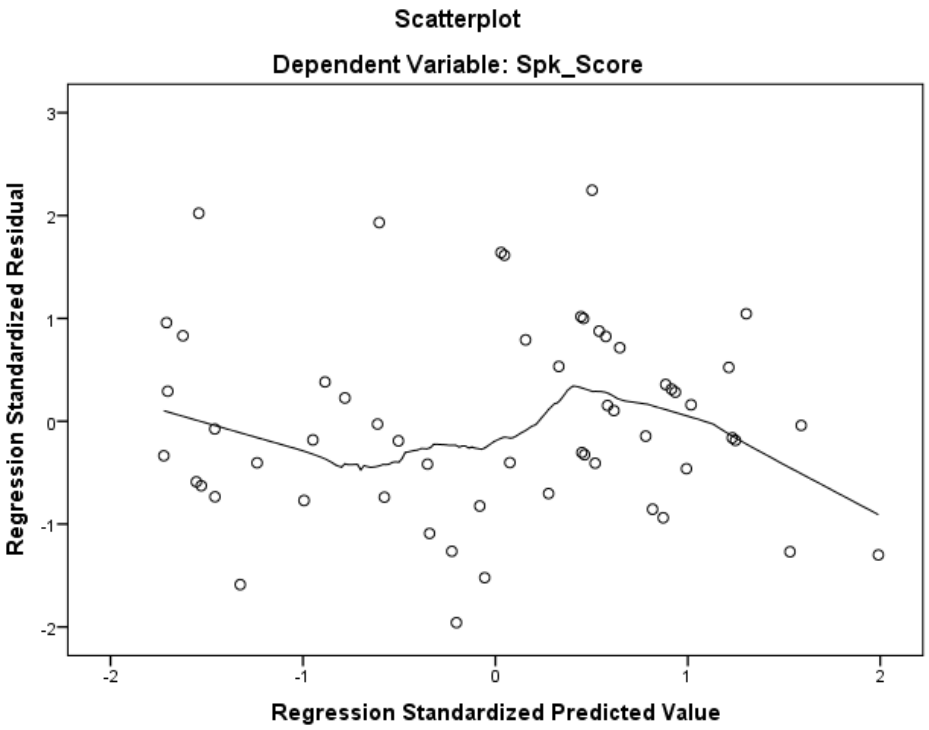


Figure 4.3. A scatterplot of standardized residuals against standardized predicted SPEAK scores

There was a significant prediction of the holistic SPEAK score by the predictor variables, $F(8, 56) = 15.738$, $p < .001$, adjusted $R^2 = .678$. Specifically, there was a significant negative prediction of SPEAK score by the proportion of disfluent collocations, $B = -74.69$, $t(56) = -6.07$, $p < .001$, $sr^2 = -.659$. All other variables did not significantly predict midterm score (see Table 4.21). The confidence limits for the proportion of disfluent collocations were -99.42 to -49.96, meaning that the decrease of SPEAK score was somewhere between -99.42 to -49.96 for an additional unit of increase in the proportion of disfluent collocations in speech while keeping the other predictor variables constant.

Table 4.20 Mean, Standard Deviation, and Intercorrelations for SPEAK Score and Predictor Variables (n=60)

Variable	Mean	SD	ACP_OK	ACP_ERR	ACP_RAT	GRA_ERR	GRA_RAT	TRAN	RES_FRE	CHOP
SPEAK Score	194.39	26.79	.62***	-.51***	.54***	-.05	.33***	.18	.39**	-.76***
ACP_OK: Acceptable collocations	12.80	5.07		-.44***	.62***	.05	.36**	.13	.63***	-.51
ACP_ERR: Unacceptable collocations	3.58	2.89			-.74***	.13	-.10	-.11	-.37**	.39***
ACP_RAT: Acceptable to unacceptable collocations	5.77	6.33				-.09	.24	.20	.51***	-.44***
GRA_ERR: Erroneous collocations	5.13	2.75					-.74***	-.09	-.31**	-.15
GRA_RAT: Error-free to erroneous collocations	3.55	3.26						.00	.50***	-.11
TRAN: Partially figurative collocations	.08	.07							.26	-.04
RES_FRE: Highly restricted collocations	4.18	2.73								-.27
CHOP: Disfluent collocations	.25	.20								

**p<.01; ***p<.00

Table 4.21 Regression Analysis Summary of a Full Model for Predicting SPEAK Score (n=60)

Variable	*B*	*CI*		*t*	*p*	sr^2
Intercept	196.79	173.11	220.46	16.71	.000	
ACP_OK: Acceptable collocations	.99	-.46	2.43	1.38	.175	.195
ACP_ERR: Unacceptable collocations	-2.27	-4.60	.047	-1.97	.055	-.273
ACP_RAT: Acceptable to unacceptable collocations	-.19	-1.32	.94	-.34	.738	-.049
GRA_ERR: Erroneous collocations	.85	-2.12	3.82	.58	.568	.083
GRA_RAT: Error-free to erroneous collocations	2.49	-.09	5.08	1.94	.058	.269
TRAN: Partially figurative collocations	59.16	-5.87	12.18	1.83	.074	.255
RES_FRE: Highly restricted collocations	-1.08	-3.29	1.13	-.98	.331	-.140
CHOP: Disfluent collocations	-74.69	-99.42	-49.96	-6.07	.000	-.659

4.3.3 The Most Parsimonious Regression Model for Predicting SPEAK Score

A stepwise forward procedure of selecting predictor variables based on order of importance was performed to find a more parsimonious model. Table 4.22 shows the statistics of R square changes across different models. It was found that Model 4 in which SPEAK score was predicted by CHOP (proportion of disfluent collocations), ACP_OK (normalized frequency of acceptable collocations), GRA_RAT (ratio of error-free to erroneous collocations), and ACP_ERR (normalized frequency of unacceptable collocations) was the most parsimonious model. The reduced model yielded an almost equivalent adjusted R^2 value (adjusted R^2 = .679) as the full model (adjusted R^2 = .678) mentioned above (see Table 4.21). Based on the reduced model, there was a significant negative prediction of

SPEAK score by CHOP (the proportion of disfluent collocations), B = -76.17, $t(56)$ = -6.38, $p < .001$, sr^2 = -.663, and by ACP_ERR (the normalized frequency of unacceptable collocations), B = -1.76, $t(56)$ = -2.19, $p < .05$, sr^2 = -.291; GRA_RAT (the ratio of error-free to erroneous collocations) significantly positively predicted SPEAK score, B = 1.50, $t(56)$ = 2.25, $p < .05$, sr^2 = .297 (see Table 4.23).

Table 4.22 Statistics for SPEAK Prediction Model Comparisons

Model (Predictors)	***Adjusted R^2***	***ΔR^2***	***ΔF***	***p***
Model 1 (CHOP)	.568	.576	74.615	.000
Model 2 (CHOP, ACP_OK)	.635	.072	11.040	.002
Model 3 (CHOP, ACP_OK, GRA_RAT)	.656	.026	4.293	.043
Model 4 (CHOP, ACP_OK, GRA_RAT, ACP_ERR)	.679	.028	4.810	.033

Table 4.23 Regression Analysis Summary of the Most Parsimonious Model for Predicting SPEAK Score (n=60)

Variable	***B***	***CI***		***t***	***p***	***sr^2***
Intercept	202.98	184.66	221.30	22.233	.000	
CHOP: Disfluent collocations	-76.17	-100.11	-52.23	-6.384	.000	-.663
ACP_OK: Acceptable collocations	.92	-.12	1.97	1.770	.083	.238
GRA_RAT: Error-free to erroneous collocations	1.50	.16	2.84	2.246	.029	.297
ACP_ERR: Unacceptable collocations	-1.764	-.38	-.15	-2.193	.033	-.291

4.3.4 Multiple Regression for Predicting TEACH Score

The aforementioned procedure of analysis was performed on the TEACH dataset as well. First, the assumptions of a multiple regression were examined. To resolve the problem of multicollinearity, GRA_OK (frequency of error-free collocations) and RES_PRO (proportion of highly restricted collocations) were removed (Table 4.24). One outlier (standardized residual > |2|) was removed from the analysis. The assumption of normality and homoscedasticity were met: Shapiro-Wilk $W = .977$, $p = .342$, Breusch-Pagan $\chi^2(8) = 6.147$, $p = .631$, respectively. Linearity assumption was also met—the scatter cloud in the scatterplot of standardized residuals (Figure 4.4) is around the horizontal line of zero. All other assumptions were met. The descriptive and correlation matrix among the variables are displayed in Table 4.25.

Table 4.24 Collinearity Statistics for the SPEAK Dataset (n=60)

Variable	**VIF**	
	Complete Model	**Model with predictors GRA_OK and RES_PRO removed**
ACP_OK	33.639*	6.420
ACP_ERR	7.740	2.914
ACP_RAT	3.198	3.096
GRA_OK	37.134*	N/A
GRA_ERR	3.294	2.951
GRA_RAT	3.048	2.780
TRAN	1.312	1.256
RES_FRE	12.668*	3.901
RES_PRO	4.641	N/A
CHOP	1.484	1.434

Note: * VIF > the cut-off value of 10

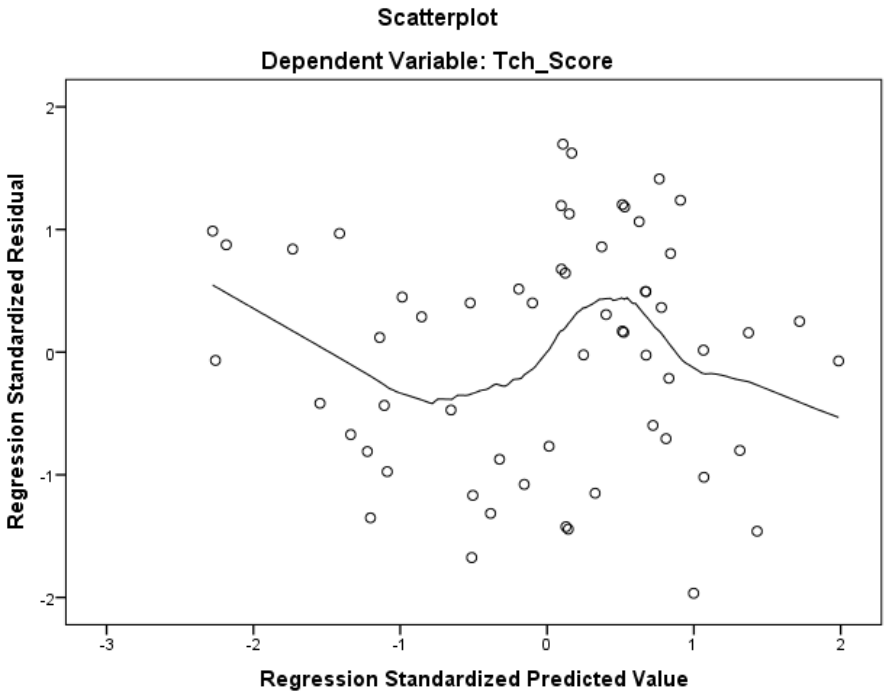


Figure 4.4. A scatterplot of standardized residuals against standardized predicted TEACH scores

In the full model, there was a significant prediction of the TEACH score by OCPMs, $F(8, 58) = 10.726$, $p < .001$, adjusted $R^2 = .573$. The significant positive predictors for TEACH score included ACP_OK (normalized frequency of acceptable collocations), $B = 2.14$, $t(58) = 2.065$, $p < .05$, $sr^2 = .280$, and GRA_RAT (the ration of error-free to erroneous collocations, $B = 1.47$, $t(58) = 2.078$, $p < .05$, $sr^2 = .282$. The significant negative predictors were ACP_ERR (the normalized frequency of unacceptable collocations), $B = -6.70$, $t(58) = -2.446$, $p < .01$, $sr^2 = -.327$, and CHOP (the proportion of disfluent collocations), $B = -84.90$, $t(58) = -6.410$, $p < .01$, $sr^2 = -.672$. All other variables did not significantly predict the TEACH score (see Table 4.25).

Table 4.25 Mean, Standard Deviation, and Intercorrelations for TEACH Score and Predictor Variables (n=60)

Variable	Mean	SD	ACP_OK	ACP_ERR	ACP_RAT	GRA_ERR	GRA_RAT	TRAN	RES_FRE	CHOP
TEACH Score	194.58	29.67	.46***	-.06	.25	.09	.32**	.16	.33**	-.68**
ACP_OK: Acceptable collocations	12.54	6.21		.12	.46***	.41**	.31**	-.01	.84***	-.40
ACP_ERR: Unacceptable collocations	2.35	2.89			-.58***	.13	.29	-.18	.03	.07
ACP_RAT: Acceptable to unacceptable collocations	6.95	6.23				.25	-.04	.21	.34**	-.33**
GRA_ERR: Erroneous collocations	1.34	1.58					-.47***	-.01	.27**	-.17
GRA_RAT: Error-free to erroneous collocations	8.75	5.99						-.05	.28	-.12
TRAN: Partially figurative collocations	.06	.07							-.12	-.37**
RES_FRE: Highly restricted collocations	7.23	4.45								-.30*
CHOP: Disfluent collocations	.25	.23								

**p<.01; ***p<.001

Table 4.26 Regression Analysis Summary of a Full Model for Predicting TEACH Score (n=60)

Variable	*B*	*CI*		*t*	*p*	sr^2
Intercept	206.52	186.35	226.68	20.572	.000	
ACP_OK: Acceptable collocations	2.14	.06	4.23	2.065	.044	.280
ACP_ERR: Unacceptable collocations	-3.68	-6.70	-.66	-2.446	.018	-.327
ACP_RAT: Acceptable to unacceptable collocations	-1.58	-3.03	-.14	-1.958	.055	-.297
GRA_ERR: Erroneous collocations	1.91	-3.65	7.46	.690	.494	.097
GRA_RAT: Error-free to erroneous collocations	1.47	.05	2.90	2.078	.049	.282
TRAN: Partially figurative collocations	-4.37	-91.78	83.04	-.100	.920	-.014
RES_FRE: Highly restricted collocations	-1.81	-4.08	.46	-1.600	.116	-.221
CHOP: Disfluent collocations	-84.90	-111.50	-58.30	-6.410	.000	-.672

4.3.5 The Most Parsimonious Regression Model for Predicting TEACH Score

A forward stepwise procedure was employed to identify the most parsimonious model for predicting the TEACH score. Model comparison statistics indicated that a regression model of the TEACH score as the dependent variable and CHOP (the proportion of disfluent collocations) and GRA_RAT (the ratio of error-free to erroneous collocations) as independent variables (Model 2 in Table 4.27) were the most parsimonious. The coefficient of multiple determination of the reduced model (adjusted R^2 = .554) is only slightly lower than that of the full model from the previous section (adjusted R^2 = .573). In the reduced model, there was a significant negative prediction of TEACH score by CHOP (the proportion of disfluent collocations), B = -81.19, $t(58)$ = -6.967, $p < .001$, sr^2 = -.678; GRA_RAT (the

ratio of error-free to erroneous collocations) was a significant predictor for TEACH score as well, $B = 1.22$, $t(58) = 2.606$, $p < .05$, $sr^2 = .326$ (see Table 37).

Table 4.27 Statistics for TEACH Prediction Model Comparisons (n=60)

Model (Predictors)	***Adjusted R^2***	***ΔR^2***	***ΔF***	***p***
Model 1 (CHOP)	.514	.522	62.361	.000
Model 2 (CHOP, GRA_RAT)	.554	.047	6.128	.016

Table 4.28 Regression Analysis Summary of the Most Parsimonious Model for TEACH Score (n=60)

Variable	***B***	***CI***		***t***	***p***	***sr^2***
Intercept	204.01	192.05	215.97	34.156	.000	
CHOP: Disfluent collocations	-81.19	-108.40	-60.00	-6.967	.000	-.678
GRA_RAT: Error-free to erroneous collocations	1.22	.282	2.154	2.606	.012	.326

4.3.6 Logistic Regression

To answer the research question on whether OCPMs predict ITA certification decisions, a simultaneous logistic regression analysis was performed on oral proficiency level as outcome and four predictors: ACP_OK (frequency of acceptable collocations), ACP_ERR (frequency of unacceptable collocations), GRA_RAT (ratio of error-free to erroneous collocations), and CHOP (proportion of disfluent collocations), the same predictor variables used by the most parsimonious models from the multiple regression analyses above. Because dichotomous certification decisions were made according to the numeric SPEAK and TEACH scores, it was assumed that these variables were also the most parsimonious predictors for certification decisions. The oral proficiency level was dichotomized into two

levels (Certified, i.e., Levels 1 and 2 and Uncertified, i.e., Levels 3 and 4). The OCPMs were calculated based on each speaker's aggregated SPEAK and TEACH responses. In other words, certification decisions were predicted by one's overall collocational performance in both SPEAK and TEACH exams.

The assumptions of logistic regression were first evaluated. Collinearity statistics indicated that multicollinearity was not a concern—none of the predictor variables had a VIF over 10. One outlier (standardized residuals exceed |2|) was removed. The assumption of linear relationships between predictor variables and a logit (boxcox tests) was also met.

Table 4.29 Logistic Regression Analysis Summary for Predicting Certification Decisions (n=60)

Variable	***B***	***SE***	***df***	***p***	***Exp(B)***
Intercept	3.51	5.50	1	.523	33.550
CHOP: Disfluent collocations	-55.95	33.94	1	.099	.000
ACP_OK: Acceptable collocations	.02	.46	1	.111	1.023
GRA_RAT: Error-free to erroneous collocations	5.39	3.58	1	.132	219.880
ACP_ERR: Unacceptable collocations	-7.44	4.67	1	.111	.001

There was a significant prediction of certification decisions by all the predictors, $\chi^2(8) = 76.34$, $p < .001$, Negelkerke $R^2 = .968$. However, there was no significant prediction of each individual predictor (see Table 4.29). There was no significant difference between observed and predicted group membership, Hosmer-Lemeshow $\chi^2(8) = .299$, p = 1.000. The overall classification rate was excellent, ROC area = .973. A cutoff value of .431 was used to minimize false negative rates (Lee, 1999). About ninety-six percent of participants were correctly classified (Certified = 96.7%; uncertified, 96.6%). Only 3.3% of high-proficiency

speakers were misclassified as low-proficiency speakers. While, 3.4% of low-proficiency speakers were misclassified as high-proficiency speakers.

4.3.7 Composite Scoring vs. Regression Models

Table 4.30 compares the effect sizes of the composite measure (CPR) and the most parsimonious regression models for predicting certification decisions and human criterion scores of oral proficiency. It is shown that the R squares (the proportion of variance in the dependent variable explained) resulting from the regression approach were generally higher than those from the composite scoring approach. This is understandable because weight assignment on all the predictor variables in the composite measure was determined by a preconceived theory (see Chapter 3, Section 3.3.4) whereas parameter estimation in the regression models was data driven.

Table 4.30 A Comparison of the Predictive Values of the Composite Scoring Approach and Regression Approach (n=60)

Oral Proficiency Measures	**CPR (R^2)**	**Parsimonious Regression Model (Negelkerke or adjusted R^2)**
Certification Decision	.681	.968
SPEAK Score	.549	.678
TEACH Score	.382	.573

4.4 Cross-validation across Two Speaking Contexts

The fourth research question investigated the context effect on the relationship between collocation use and perceived oral proficiency. It was hypothesized that collocation contributes to oral proficiency at least differently in different contexts of language use. To

test this hypothesis, the best fitted regression model obtained from the SPEAK dataset was used to predict oral proficiency scores in the TEACH dataset (i.e., cross-validation).

First, the intact regression model for predicting SPEAK scores in Section 4.4.1 were used to compute predicted TEACH scores. In other words, the SPEAK dataset were considered a training sample for model construction. Then, the correlation between the predicted TEACH scores and observed TEACH scores were calculated. The square of this correlation coefficient was the cross-validation R square. In this case, the correlation coefficient was $r = .742$ ($p < .01$) and the cross-validation R square was .551, which was less than the R square, $R^2 = .678$, in the SPEAK dataset. The R square difference across two datasets indicates a shrinkage in predictive power when the regression model trained on the SPEAK exam data was used to predict oral proficiency scores in the TEACH exam. This suggests that the same prediction model cannot be applied to two significantly different speaking tasks.

4.5 Chapter Summary

This chapter examined the empirical performance of eleven multi-dimensional operational collocational performance measures (OCPMs) for predicting oral English proficiency determined by human raters. It was found that a majority of OCPMs alone had moderate to strong predictive values. When these predictor variables were used to construct a regression prediction model, the predictive values further increased. Particularly, the collocation-based regression model had nearly a perfect prediction on ITA certification decisions, thus lending support for the theory that spoken collocational competence (SCC) is a core component of L2 oral construct. Construction of parsimonious models was also

attempted. It was found that two to four OCPMs as predictor variables yielded similar predictive power as the full model that included all OCPMs. In addition, the composite scoring approach (CPR) inspired by sports performance rating and the regression approach were compared; the latter approach was found to predict certification decisions and holistic rater scores better. Finally, the hypothesis that regression model specification should change for assessing oral proficiency in different contexts was tested using a cross-validation technique. Evidence was obtained to support this hypothesis.

CHAPTER 5: CONCLUSIONS AND DISCUSSION

This chapter begins with a summary and discussion of the major findings on the four research questions. It then proceeds with the implications for L2 speaking theory, automated speech evaluation, and training Chinese ITAs. The chapter ends with a discussion of the limitations of this study and directions for future research.

5.1 A Summary and Discussion of the Major Findings

This section summarizes the major findings on the four research questions raised in Chapter 2: the basic characteristics of the collocation occurrences in learner speech, the differences in the collocation measures among proficiency level groups, the prediction of L2 oral proficiency based on the collocation measures, and the context effect on the prediction.

5.1.1 Basic Characteristics of Collocations

The first research question is "What are basic characteristics of the collocations that ESL learners produce in naturally occurring speech?" To answer this question, descriptive statistics from collocation coding were examined. In this study, the following characteristics of collocation use were investigated: frequency, accuracy, complexity and fluency.

5.1.1.1 Frequency

It was found that the sixty Chinese ITAs used collocations frequently in English speaking. Specifically, a learner, on average, produced 20.42 and 16.70 lexical collocations per 500 words in the SPEAK and TEACH exams, respectively. Of the ten targeted syntactic patterns, the Chinese ITAs used verb-noun (44%), adjective-noun (25%), and noun-noun

(10%) collocations most frequently but the remaining patterns infrequently. These numbers are comparable to the findings from previous research. Sung (2003) reported an average frequency of 13.15 lexical collocations per 400 words in seventy-two college ESL learners' oral remarks on a film. Xu and Xi (2010) found that verb-noun, adjective-noun, and noun-noun collocations comprised a majority (84.7%) of the lexical collocations in 429 human transcriptions of TOEFL iBT Speaking Practice test responses.

Chinese ITAs' heavy reliance on verb-noun, adjective-noun, and noun-noun collocations may be attributed to the fact that these patterns always convey important content information in oral communication. Compared to them, the other patterns may not be as indispensable. For example, amplifier collocations, including adverb-adjective (e.g., highly accurate, deeply disappointed) and adverb-verb collocations (e.g. whisper softly), contain an optional intensifier. They may be replaced by single verbs (e.g., whisper) or free combinations (e.g., very accurate, disappointed a lot) in informal speech. Unfortunately, no relevant empirical research can be found to support this speculation. Some indirect evidence for the explanation comes from Molavi, Koosha, and Hosseini's (2014) content analysis on three English language teaching (ELT) textbooks written by native authors. It was found that verb-noun and adjective-noun collocations occurred far more frequently than other collocation patterns (noun-verb, noun-of-noun, adverb-adjective, and verb-adverb) in these textbooks.

An equally plausible explanation for the finding concerns syntactic transfer (also called formal transfer) in which L2 learners map L2 lexical items onto an L1 syntactic structure. Kormos (2006) posits that syntactic transfer can reduce the cognitive intensity in L2 speaking as it saves mental resources that would otherwise be needed for grammatical

encoding (see Chapter 2, Sections 2.3.2 and 2.3.4). Given the large number of verb-noun, adjective-noun, and noun-noun collocations in Mandarin Chinese (see Mei, 1999), the Chinese learners of English may have felt tempted and intuitive to draw on these existing language resources to facilitate L2 speaking.

5.1.1.2 Accuracy

This study found that the Chinese ITAs generally tended to produce fewer unacceptable collocations ($Z = -2.968$, $p < .01$) and ungrammatical collocations ($Z = -6.331$, $p < .01$) in the TEACH exam than in the SPEAK exam. The observed differences in collocation accuracy, in my opinion, may be explained by two sources of variance based on speech-processing theory, one external to the examinee and the other internal to the examinee.

The external source of variance concerns the testing method. Specifically, the SPEAK exam elicited completely impromptu speech whereas the TEACH exam elicited somewhat prepared speech—although one hour was not adequate for preparing a well-rehearsed lecture, it gave the examinees an opportunity for conceptual planning in advance. In addition, the SPEAK exam involved some language behaviors unseen in the TEACH exam (the solo presentation component), such as listening comprehension, turn-taking, and meaning negotiation with an interlocutor. Overall, the SPEAK was more cognitively demanding than the TEACH exam; the former engaged some extra cognitive and affective processes (e.g., speech decoding, recognizing cues for turn-taking, analyzing the interlocutor's prior knowledge of the topic, instantaneous conceptual preparation from scratch, and strong feelings of anxiety and nervousness) that assumingly took away enormous attentional

resources from lexico-grammatical encoding and self-monitoring and, in turn, reduced the accuracy of overt speech.

The internal source of variance pertains to examinees' Spoken Collocational Competence (SCC). As discussed in Chapter 2, Section 2.4, the construct of SCC is considered context-variant. That is, the facet of SCC engaged by a specific domain of language use is defined as the overall strengths and density of the collocational connections among the most frequently used lexical items in that domain. The Chinese ITAs in this study might have higher level of SCC in the academic domain than in the social domain owing to their prior language learning experience. While they had received substantial exposure to academic English, particularly the English language used in their areas of study, they had had very limited experience of conversing with native English speakers on daily-life topics. This could explain why their collocation use in the academic domain tended to be less error-prone.

As above, the Chinese ITAs higher collocation accuracy in the academic domain seemed to be a result of the interplay between the external contextual variables and the internal learner attribute, namely SCC. The finding lends support for conceptualizing the construct of SCC in an interactionalist construct framework (Bachman, 1990, 2007; Chapelle, 1998).

5.1.1.3 Complexity

This study measured collocation complexity in two directions: restrictedness (precision) and transparency. It was found that the Chinese ITAs used highly restricted collocations more frequently in the TEACH exam (42.5%) than in the SPEAK exam (19.7%). The finding is self-explanatory in that academic language, which is used to

disseminate scientific ideas, is usually more concise, precise, and authoritative than language of daily use (Snow, 2010).

On the other hand, the Chinese ITAs produced a very small portion of partially figurative collocations (e.g., throw a party) in the SPEAK (7.5%) and TEACH (6.4%) exams. This seems to indicate their limited knowledge of idiomatic language in English. Learning idiomatic language is considered extremely difficult for L2 learners who are not residing in an English-speaking country because acquiring idiomatic expressions usually demands hearing them in comprehensible contexts and/or co-constructing the meanings of them with native speakers (Bortfeld & Brennan, 1997; Crutchley, 2007).

5.1.1.4 Fluency

The Chinese ITAs were found to pronounce a majority of the collocations smoothly in both SPEAK (75.9%) and TEACH (79.8%) exam responses. The finding provides empirical support for the hypothesis that, like native speakers, ESL learners also rely on formulaic language to maintain oral fluency (Wood, 2010; Wray & Fitzpatrick, 2010).

5.1.2 Differences among Proficiency Level Groups

The second research question is "What collocation measures can effectively differentiate among ESL speakers at different oral proficiency levels?" To answer this question, one-way analysis of variance (ANOVA) was performed on each operational collocation performance measure as dependent variables and proficiency level group as an independent variable. The ANOVA results indicated that most collocation measures met the theoretical expectation. Specifically, all of the measures, except transparency, were able to

differentiate between a certified group (Level 1 and Level 2) and an uncertified group (Level 3 and Level 4) in a meaningful way. The transparency measure did not perform as expected mainly because partially figurative collocations were rare in learner speech (see the previous section).

A few measures even differentiated between two adjacent proficiency level groups. These included GRA_ERR or the frequency of erroneous collocations (Level 1 < Level 2, $p < .01$), GRA_RAT or the ratio of error-free to erroneous collocations (Level 1 > Level 2, $p < .01$), CHOP or disfluencies in collocation utterances (Level 3 < Level 4, $p < .01$), and CPR or the composite measure that adds up positive collocational features and deducts negative ones (Level 3 > Level 4, $p < .01$).

A finding was, however, counterintuitive at first look. It was found that Level 4 speakers, the lowest proficiency group, on average produced erroneous collocations less frequently than Level 3 speakers ($p < .01$). However, we cannot interpret this finding in isolation without taking into consideration three additional findings. First, a Level 4 speaker, on average, produced six fewer collocations than a Level 3 speaker. Second, a Level 4 speaker produced a larger proportion of disfluent collocations than a Level 3 speaker ($p < .01$). Third, Level 4 speakers' overall collocational performance as measured by the composite measure (CPR) was significantly lower than Level 3 speakers ($p < .01$).

As discussed in Chapter 3, test-taker behaviors may be affected by two types of motivation: *promotion focus* on seeking accomplishments and *prevention focus* on seeking safety (Halvorson & Higgins, 2013; Higgins, 1997). Xu and Xi (2010), for example, noted that the lowest-level speakers in their sample tended to play safe and avoid using complicated collocations. In this study, the lowest-level speakers seemed to play safe by focusing on

reducing minor surface errors. They did achieve somewhat high grammatical accuracy, however, at the cost of producing fewer collocations. Still, their overall collocational performance ranked the lowest among the four proficiency groups. This reminds us that L2 learners' oral collocational performance is multi-dimensional and must be evaluated as a whole.

5.1.3 Prediction for L2 Oral Proficiency

The third research question is "To what extent can multidimensional collocation measurement predict human judgement of L2 oral proficiency?" To answer this question, correlation and regression analyses were conducted. The results of Pearson correlations between OCPMs and human criterion scores suggested that the Chinese ITAs' collocational performance in naturally occurring speech was in many ways related to the holistic human scores of oral proficiency. The measures on semantic accuracy, grammatical accuracy, and restrictedness showed weak to moderate relationships with the human scores (r = .311-.497, p < .01 in SPEAK; r = .331-.457, p < .01 in TEACH). The measure on automaticity had a moderate to strong relationship with the human scores (r = -.714, p < .01 in SPEAK; r = -.675, p < .01 in TEACH). At the same time, the collocation measures had no relationship with speech length, thus discounting a potential counterargument that the relationships between the collocation measures and human scores were mediated by speech length.

The magnitudes of the relationships between semantic accuracy of collocation use and human scores were comparable to previous research. Sung (2003) found such relationships to be weak to moderate (r = -.320-.545, p < .01). Xu and Xi (2010) found the relationships to be weak (r = .271-.380, p < .01). However, the spoken responses in their

study were relatively short so the observed relationships could have been attenuated by measurement errors due to insufficient observation of collocation use (Thorndike, 1951).

The only single collocation measure that was found to strongly relate to human scores was CHOP or the disfluencies of collocation utterance in the SPEAK exam. This observed strong relationship is interpretable based on theory discussed in Chapter 2. That is, producing lexical collocations as phonologically coherent units is supposed to create speech rhythm and increase oral fluency. In the literature, fluency or temporal measures on articulation rate, the length of runs, pauses, and silences have been frequently reported as weak to moderate predictors for holistic scores of oral proficiency (e.g., Ginther, Dimova, & Yang, 2010; Xi, et al., 2008). This is probably because human raters are sensitive to the disfluencies in learner speech (see Brown, et al., 2005, p. 24). However, the interpretations of the temporal measures are not as straightforward as the present CHOP measure. An evaluation of speech characteristics without looking at speech content, in my opinion, is coarse. Collocation, as found in this study, seems to be the exact construct to make this connection between rhythm and meaning and also between the speaking process and speech product.

The multiple regression and logistic regression analyses, on the other hand, suggested that the optimal effect sizes of multi-dimensional collocation measurement for predicting human holistic scores were approximately 68% (adjusted R^2 = .678) in the SPEAK exam and 57% (adjusted R^2 = .573) in the TEACH exam; the effect size for predicting certification decisions (certified or not certified) was nearly 97% (Negelkerke R^2 = .968). That is to say, disregarding the Chinese ITAs' other aspects of speaking performance (e.g., pronunciation, sentence-level grammar, use of cohesive devices, etc.), their collocational performance in speaking alone account for above 50% of the variance in their human criterion scores of oral

proficiency. The results provide empirical support for my argument that collocation measurement is the key to assessing L2 oral proficiency (Chapter 2, Section 2.4).

5.1.4 Context Effect on Prediction

The fourth research question is "Would the magnitude of the relationship between collocation and L2 oral proficiency vary across two distinct speaking contexts?" To answer this question, a cross-validation technique based on regression was performed between the SPEAK and TEACH datasets. It was found that the multiple regression model found to best predict human holistic scores in the SPEAK exam lost substantial amount of predictive power (12.7%) when used to predict human holistic scores in the TEACH exam. It is worth mentioning that raters were not a source of variance in this analysis as each examinee was scored by the same panel of human raters across the two exams.

This finding suggests that an L2 speaker's collocational performance may be weighed differently in human perception of oral proficiency in the two exams. Stated another way, the L2 oral constructs from the two contexts of language use (socialization and teaching in a narrow academic domain) are somewhat different in their composition. The social oral construct seems to comprise more linguistic factors (e.g., spoken collocational competence) than the academic oral construct. Schmidgall (2013), for example, found that listeners' impression of an ITA's oral language ability for teaching was influenced by many non-linguistic factors such as the speaker's personality and teaching effectiveness (e.g., using the chalkboard and nonverbal communication) and the listeners' background knowledge and interest in the topic. Thus, although the Chinese ITAs, as found in this study, used lexical

collocations fairly proficiently in their field of study, this distinguished performance may not have been awarded by human raters.

5.2 Implications

The implications of the major findings from this study have both theoretical and practical implications for L2 speaking theory, automated speech evaluation, and language teaching.

5.2.1 Implications for L2 Speaking Theory

Although there is a growing consensus in the second language acquisition (SLA) literature that collocational competence is a defining aspect of language proficiency (Schmitt, 2010), contemporary L2 speaking theories (e.g., Bygate, 1987; Kormos, 2006) fall short of giving an adequate explanation of the role that collocation plays in speech formulation. This could be because little empirical research has shown that collocation in learner language directly enhances speaking performance (Millar, 2011).

This study formulated a new construct called SCC based on a logical analysis of the connections among collocation, speech-processing theories, and rubrics for oral language assessment. In light of Skehan's (1998, 2009) trade-off hypothesis, the study also developed a number of operational measures for SCC and investigated their empirical performance with real L2 oral assessment data. The finding that the collocation measures, as a whole, explained a large proportion of the variance in human criterion scores of oral proficiency has lent support for the working theory that proceduralization of oral collocation production offsets

the cognitive load involved in lexical selection and thus contributes enormously to L2 oral proficiency (Chapter 2, Section 2.4).

5.2.2 Implications for Automated Speech Evaluation

Some language testing researchers maintain that assessing speaking skills is distinct from assessing other language skills in that the former elicit learner language both as a process and a product (Fulcher, 2003; Luoma, 2004). L2 spoken data therefore contain rich information that may reflect the underlying processes of speech formulation.

In the common practices of L2 speaking assessment, a test taker's collocational performance is usually not specifically rated mainly because human raters can only focus on a limited range of speech characteristics in live rating. However, automated scoring technologies give us high hope to look into L2 learners' oral language behaviors in more detail.

The findings of this study suggest that L2 learners' collocational performance in free speech deserve examiners' closer attention. The collocations in learner speech (particularly, the most frequently used verb-noun, adjective-noun, and noun-noun collocations) seem to contain useful information for predicting human judgment on oral proficiency. As construct underrepresentation is a major drawback in automated speech evaluation of today, it is recommended that NLP researchers automate the collocation measures investigated in this study. Once automated speech recognition technology permits accurate L2 speech transcription, these collocation measures will be readily applicable.

Although this study found that the most parsimonious prediction model for oral proficiency only contained two to four collocation measures as predictor variables, it is not

recommended that NLP researchers only target these measures. As this paper repeatedly emphasized, collocation measurement is multi-dimensional and should be performed as a whole. On the one hand, L2 speakers, as found in this study, may sacrifice one aspect of collocational performance (e.g., automaticity) for another one (e.g., grammatical accuracy). It was a speaker's overall collocational performance in speaking that predicted his or her perceived oral proficiency. On the other hand, the goal of automated speech evaluation is not limited to providing a holistic test score. As mentioned in Chapter 2, a great advantage of automated scoring to human scoring is its potential for providing detailed individualized feedback to language learners. Hence, the three dimensions of collocation measurement (accuracy, complexity, and fluency) are equally important for detailed feedback generation.

5.2.3 Implications for Training Chinese ITAs

At Iowa State University, Chinese graduate students comprise a large body of ITAs (48%) and are heavily relied on to perform various teaching duties on campus. At the same time, they are the largest student body enrolled in ITA remedial classes. It is in both the university and these students' interest that they pass the SPEAK and TEACH exams and become eligible for teaching.

This study has shown that Chinese ITAs' collocation performance in the two exams moderately predicted their resultant test scores and certification outcomes. The finding implies that improving low-level speakers' SCC may help improve their oral proficiency and increase their chances of passing the exams.

This study has identified four issues inherent in Chinese ITAs oral collocation usage. First, it was found that Chinese ITAs avoided adverb-adjective (0.4%), adverb-verb (3.7%),

and noun-verb (1.9%) collocations in speaking. This may indicate their lack of knowledge of these patterns. Second, Chinese ITAs were found to use a large number of verb-noun and phrasal-verb-and-noun collocations in speech but their accuracy on these two patterns was relatively low (64.8% and 57.0%). Previous research has also revealed Chinese ITAs' difficulty with these two patterns (e.g., Liao & Fukuya, 2004; Voss, 2012). Their difficulty with phrasal verbs may be ascribed to the L1-L2 structural differences. That is, unlike English phrasal verbs, Chinese phrasal verbs are inseparable and rarely convey figurative meanings (Liao & Fukuya, 2004). The deviant verb-noun collocations may likely be caused by unsuccessful semantic transfer (Kormos, 2006), i.e., translating Chinese collocations directly into English (e.g., from '学知识' to 'learn knowledge'). Third, this study found that the Chinese ITAs barely used partially figurative collocations which are, however, commonly observed in native English speech. Fourth, the Chinese ITAs seemed to use collocations more proficiently in the academic domain than in the social domain probably owing to their limited exposure to the latter.

For the reasons above, it is recommended that the instructors of the remedial courses allocate adequate teaching resources to help Chinese ITAs improve the above aspects of collocation performance in English speaking. SLA researchers have found the following techniques useful:

1. Drawing learners' attention to collocations, particularly those that are markedly different in form from their native-language (L1) equivalents (e.g., Laufer, 2011; Wood, 2009; Zughoul & Abdul-Fattah, 2003).
2. Providing learners with a list of useful collocations that they can rely on for participating in task-based language learning activities (Wray & Fitzpatrick, 2010).

3. Increasing learners' exposure to collocations in meaningful contexts (Durrant, 2008; Myers & Chang, 2009).
4. Relating target collocations to learners' prior knowledge (Yasuda, 2010).
5. Having learners analyze their own collocation production before offering corrections (Nesselhauf, 2003).
6. Teaching the collocations that share the same node word (e.g., obtain/gain/acquire knowledge) together but avoid any pairs of synonyms (Webb & Kagimoto, 2011).

5.3 Limitations and Directions for Future Research

This dissertation study is limited in three ways. First, the study was performed on a relatively small sample (n = 60) in consideration of feasibility—human speech transcription and manual collocation identification and coding were extremely time-consuming and costly. Because of the small sample size, structural equation modelling (SEM), a more advanced quantitative research method, was not employed to investigate the relationships among collocation measures, human holistic scores, and human analytical scores in the two speaking contexts. This investigation would reveal how collocation usage in speech contributes to the various aspects of observed speaking performance that human raters are trained to focus on in determining holistic scores. Also because only a small sample was affordable, the participants' native language was controlled. Considering that L2 collocation use may interact with learners' native language (Nesselhauf, 2005), the results of this study are not generalizable to other L1 groups.

The second limitation of the study concerns the unsatisfactory coder agreement on semantic accuracy (collocation acceptability). In this study, human coding on this feature

resulted in a considerable number of discrepancies (560 out of 2344) and a relatively low Kappa statistic (k = .254). This was a familiar scene from Xu and Xi's (2010) study which also found native coders' agreement on this feature to be low (k = .180). These findings call into question the use of human judgment as a gold criterion for judging collocation acceptability.

Based on my communication with the coders, I realized that their collocational knowledge were somewhat limited in certain domains of language use. In the following excerpt, a coder expressed his concern on coding collocations from a field he was not familiar with.

> I'm not familiar with this as a collocation, but it sounds like it could be a familiar word combination in the biological sciences. What to do in this case? (Coder 12)

Native speakers are often treated as experts of a language. However, it seems untrue that they are experts in all domains of language use. For example, collocations like 'a sweet spot' and 'a dry spot' are common collocations in basketball language indicating certain regions on the court where an athlete has high shooting percentage and low shooting percentage. These collocations may not be familiar to native speakers who are laypersons of basketball.

As discussed in Chapter 2, the collocational network in a language learner's mental lexicon is assumingly developed based on his or her prior language use experience (e.g., life experience, reading, education, ESL teaching experience, exposure to world Englishes, etc.). Thus, even native speakers' collocational knowledge can be unbalanced across specific

topics or domains of language use. The native-English-speaking coders used in this study comprise undergraduate students, graduate students, full-time ESL instructors, and senior faculty members in an Applied Linguistics program. It is unsurprising that the breadth and depth of their collocational knowledge vary and that they disagree on the acceptability of certain L2 collocations.

In this study, a remedy taken was drawing on native speaker's collective knowledge and having the discrepancies resolved by a third coder. However, human knowledge of collocations is always limited. It is recommended that future studies base acceptability judgment on large reference corpora of native speech. Corpora can be a resource for identifying typical collocations in the target language domain.

Third, there was no qualitative component in this research study. That is, the relationship between collocation and L2 oral proficiency was analyzed in the light of contemporary speech-processing theories and examined objectively with real data from a third person's perspective. Qualitative data such as raters' verbal reports on their rating orientations (e.g., Brown, et al., 2005) would provide support to the interpretations of the quantitative research findings. Unfortunately, the spoken data and test results were masked secondary data from SPEAK and TEACH exams administered between 2006 and 2011. Therefore, it was unfeasible to interview the raters and the L2 speakers from the past. It is thus suggested that future research investigate human raters' perception and L2 speakers' self-reflections on collocation use in L2 spontaneous speaking or perform functional analysis of the collocation occurrences in the transcripts (Halliday, 2013). Such qualitative research would complement the quantitative research findings of this study.

5.4 Concluding Remarks

Training and administration of human raters for L2 oral exams are expensive and highly demanding and inconsistent and construct-irrelevant rater behaviors can undermine the plausibility of test score interpretation and use (Brown, 2012). The pursuit of automated speech evaluation reflects language testers' good wish to make L2 oral exams more efficient, cost-effective, reliable, and informative to language teaching and learning. However, this pursuit also burdens the language testing community with the responsibility of defining the L2 oral construct in meticulous detail so that construct-relevant scoring features can be developed accordingly. Unfortunately, our understanding of the L2 oral construct in terms of what it is composed of and how it interacts with specific contexts of language use is limited (Chapelle, et al., 2008). This limitation in knowledge largely restricts NLP researchers' vision for designing high-level scoring features for constructed L2 oral responses.

This study has set a good example of drawing on construct theory to develop potentially useful high-level scoring features for automated speech evaluation. It explored the empirical relationship between collocation and L2 oral proficiency and provided practical guidelines for measuring the collocation occurrences in spontaneous learner speech. However, this dissertation, which only used a holistic human score as a criterion measure of oral proficiency and focused on a single L1 (Chinese) group, is limited and thus merely a starting point. More foundational research that builds upon the current one and explores the meaning of L2 oral construct in different contexts of language use (e.g., Schmidgall, 2013) is needed to move the field of automated scoring forward.

REFERENCES

APA. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*(2, supplement), 1-38.

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Alamar, B. C. (2013). *Sports analytics: A guide for coaches, managers, and other decision makers*. New York: Columbia University Press.

Al-Zahrani, M. S. (1998). *Knowledge of English lexical collocations among male Saudi college students majoring in English at a Saudi university* (Unpublished doctoral dissertation). Indiana University of Pennsylvania, Indiana, Pennsylvania.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly, 25*(4), 671-704.

Bachman, L. F. (2006). Generalizability: A journey into the nature of empirical research in applied linguistics. In M. Chalhoub-Deville, C. A. Chapelle & P. Duff (Eds.), *Inference and generalizability in applied linguistics: Multipe perspectives* (pp. 165-207). Amsterdam: John Benjamins.

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. D. Fox (Ed.), *Language testing reconsidered* (pp. 41-71). Ottawa: University of Ottawa Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Bahns, J. (1993). Lexical collocations: A contrastive view. *ELT Journal, 47*(1), 56–63.

Barnbrook, G. (2007). Sinclair on collocation. *International Journal of Corpus Linguistics, 12*(2), 183-199.

Bednarek, M. (2008). Semantic preference and semantic prosody re-examined. *Corpus Linguistics and Linguistic Theory, 4*(2), 119-139.

Bennett, R. E. (2004). *Moving the field forward: Some thoughts on validity and automated scoring* (ETS Research Rep. No. RM-04-01). Princeton, NJ: Educational Testing Service.

Bennett, R. E., & Bejar, I. I. (1998). Validity and automad scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17*(4), 9-17.

Benson, M., Benson, E., & Ilson, R. (1986). *Lexicographic description of English*. Philadelphia: John Benjamins.

Benson, M., Benson, E., & Ilson, R. (1997). *The BBI dictionary of English word combinations* (Rev. ed.). Amsterdam; Philadelphia: John Benjamins.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27*(3), 355-377.

Biber, D., Conrad, S., & Leech, G. (2002). *Longman student grammar of spoken and written English*. Harlow, Essex: Longman.

Bolinger, D. (1976). Meaning and memory. *Forum Linguisticum, 1*(1), 1-14.

Bonk, W. J. (2000). *Testing ESL learners' knowledge of collocations*. (ERIC Document Reproduction Service No. ED 442 309). Retrieved from http://files.eric.ed.gov/fulltext/ED442309.pdf

Bortfeld, H., & Brennan, S. E. (1997). Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Processes, 23*(2), 119-148.

Brown, A. (2012). Interlocutor and rater training. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 413-425). London: Routledge.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An Examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks*. Princeton, NJ: Educational Testing Service.

Bygate, M. (1987). *Speaking*. Oxford: Oxford University Press.

Carnegie Speech. (2015). Carnegie Speech Assessment™. Retrieved from https://www.carnegiespeech.com/sales/index.php/vstore/overview

Carter, R. (1998). *Vocabulary: Applied linguistic perspectives*. New York: Routledge.

Celce-Murcia, M., Brinton, D., & Goodwin, J. M. (1996). *Teaching pronunciation: A reference for teachers of English to speakers of other languages*. Cambridge: Cambridge University Press.

Chalhoub-Deville, M. (1995). A contextualized approach to describing oral language proficiency. *Language Learning, 45*(2), 251-281.

Chalhoub-Deville, M. (2003). Second language interaction: current perspectives and future trends. *Language Testing, 20*(4), 369-383.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*(4), 293-332.

Chang, H.-H. (2012). Making computerized adaptive testing diagnostic tools for schools. In R. W. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessments: Recent history and predictions for the future* (pp. 195-226). Charlotte, NC: Information Age.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-70). Cambridge, UK: Cambridge University Press.

Chapelle, C. A. (2012a). Conceptions of validity. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 21-33). London: Routledge.

Chapelle, C. A. (2012b). Validity argument for language assessment: The framework is simple... *Language Testing, 29*(1), 19-27.

Chapelle, C. A., & Chung, Y.-R. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing, 27*(3), 301-315.

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge: Cambridge University Press.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 1-25). New York: Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3-13.

Chun, C. W. (2006). An analysis of a language test for employment: The authenticity of the PhonePass test. *Language Assessment Quarterly, 3*(3), 295-306.

Chun, C. W. (2008). Comments on 'Evaluation of the usefulness of the Versant for Englsih Test: A response': The author responds. *Language Assessment Quarterly, 5*(2), 168-172.

Clauser, B. E., Kane, M. T., & Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education, 15*(4), 413-432.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213-220.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics, 29*(1), 72-89.

Cowie, A. P. (1978). The place of illustrative material and collocations in the design of a learner's dictionary. In P. Strevens (Ed.), *In honour of A.S. Hornby* (pp. 127-139). Oxford: Oxford University Press.

Cowie, A. P. (1981). The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics, 2*(3), 223-235.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oaks: Sage Publications.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Crutchley, A. (2007). Comprehension of idiomatic verb + particle constructions in 6- to 11-year-old children. *First Language, 27*(3), 203-226.

Davies, M. (2012) *The Corpus of Contemporary American English: 450 million words, 1990-2012*. Available online at http://corpus.byu.edu/coca/

de Bot, K. (1992). A bilingual production model: Levelt's "speaking" model adapted. *Applied Linguistics, 13*(1), 1-24.

de Bot, K. (2004). The multilingual lexicon: modeling selection and control. *International Journal of* Multilingualism, *1*(1), 17-32.

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*(4), 227-268.

DeKeyser, R. (2007). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 97-114). Mahwah, NJ: Lawrence Erlbaum Associates.

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition, 31*(4), 533-557.

Douglas, D. (1998). Testing methods in context-based second language research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 141-155). Cambridge, UK: Cambridge University Press.

Douglas, D. (2000). *Assessing Languages for Specific Purposes*. Cambridge: Cambridge University Press.

Douglas, D., & Selinker, L. (1992). Analyzing oral proficiency test performance in general and specific purpose contexts. *System, 20*(3), 317-328.

Downey, R., Farhady, H., Present-Thomas, R., Suzuki, M., & Van Moere, A. (2008). Evaluation of the usefulness of the Versant for English test: A response. *Language Assessment Quarterly*, *5*(2), 160-167.

Du Bois, J. W. (1991). Transcription design principles for spoken discourse research. *Pragmatics, 1*(1), 71-106.

Du Bois, J. W., Schuetze-Coburn, S., Cumming, S., & Paolino, D. (1993). Outline of discourse transcription. In J. A. Edwards & M. D. Lampert (Eds.), *Talking data: transcription and coding in discourse research* (pp. 45-89). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Durrant, P. (2008). *High frequency collocations and second language learning* (Unpublished doctoral dissertation). University of Nottingham, Nottingham, United Kingdom.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes, 28*, 157-169.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics, 47*(2), 157-177.

Durrant, P., & Schmitt, N. (2010). Adult learners' retention of collocations from exposure. *Second Language Research, 26*(2), 163-188.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition, 18*(1), 91-126.

Ellis, N. C. (2008). Phraseology: The periphery and the heart of language. In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 1-14). Amsterdam: John Benjamins.

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford University Press.

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing, 27*(3), 317-334.

Evanini, K., Xie, S., & Zechner, K. (2013). Prompt-based content scoring for automated spoken language assessment. *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, 157-162.

Evert, S. (2005). *The statistics of word cooccurrences: Word pairs and collocations*. (Unpublished doctoral dissertation). University of Stuttgart, Stuttgart, Germany.

Fan, J. (2014). Chinese test takers' attitudes towards the Versant English Test: A mixed-methods approach. *Language Testing in Asia, 4*(6), 1-17.

Fernando, C. (1996). *Idioms and idiomaticity*. Oxford: Oxford University Press.

Firth, J. R. (1968). A synopsis of linguistic theory, 1930-55. In F. R. Palmer (Ed.), *Selected papers of J. R. Firth 1952-59* (pp. 168-205). Bloomington: Indiana University Press.

Fitzpatrick, T. (2006). Habits and rabbits: Word associations and the L2 lexicon. *EUROSLA Yearbook, 6*(1), 121-146.

Fitzpatrick, T. (2007). Word association patterns: unpacking the assumptions. *International Journal of Applied Linguistics, 17*(3), 319-331.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378-382.

Fulcher, G. (2003). *Testing second language speaking*. New York: Pearson Longman.

Futagi, Y., Deane, P., Chodorow, M., & Tetreault, J. (2008). A computational approach to detecting collocation errors in the writing of non-native speakers of English. *Computer Assisted Language Learning, 21*(4), 353-367.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences, 8*(1), 8-11.

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing, 27*(3), 379-399.

Gitsaki, C. (1999). *Second Language Lexical Acquisition: A Study of the Development of Collocational Knowledge*. San Francisco: International Scholars Publications.

Glucksberg, S. (1993). Idiom meanings and allusional content. In C. Cacciari & P. Tabossi (Eds.), *Idioms: Processing, structure, and interpretation* (pp. 3-26). Hillsdale, NJ: Lawrence Erlbaum Associates.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (Ed.), *Phraseology: theory, analysis, and applications* (pp. 145-160). Oxford: Oxford University Press.

Gries, S. T. (2008). Phraseology and linguistic theory: A brief survey. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 3-25). Amsterdam: John Benjamins.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 65-110). Westport, CT: Praeger.

Halleck, G. B., & Moder, C. L. (1995). Testing language and teaching skills of international teaching assistants: The limits of compensatory strategies. *TESOL Quarterly, 29*(4), 733-758.

Moder, C. L., & Halleck, G. B. (2009). Planes, politics and oral proficiency: Testing international air traffic controllers. *Australian Review of Applied Linguistics, 32*(3), 25.21-25.16.

Moder, C. L., & Halleck, G. B. (2012). Designing language tests for specific social uses. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 137-149). London: Routledge.

Halliday, M. A. K. (2013). *Halliday's Introduction to Functional Grammar*. London: Routledge.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Halvorson, H. G., & Higgins, E. T. (2013). *Focus: Use different ways of seeing the world for success and influence*. New York: Penguin Group.

Handl, S. (2008). Essential collocations for learner of English: The role of collocational direction and weight. In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 43-66). Amsterdam: John Benjamins.

Hartsuiker, R. J., Pickering, M. J., & Veltkamp, E. (2004). Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish-English bilinguals. *Psychological Science, 15*(6), 409-414.

Hauck, M. C., Wolf, M. K., & Mislevy, R. (2013). *Creating a next-generation system of K–12 English learner (EL) language proficiency assessments.* Retrieved from https://www.ets.org/s/research/pdf/24473_K12_EL_Paper.pdf

Herbst, T. (1996). What are collocations: Sandy beaches or false teeth? *English Studies, 77*(4), 379-393.

Higgins, E. T. (1997). Beyond pleasure and pain. *The American Psychologist, 52*(12), 1280-1300.

Hollinger, J. (2004). *Pro basketball forecast*. Washington, D.C.: Brassey's Inc.

Howarth, P. A. (1996). *Phraseology in English academic writing: Some implications for language learning and dictionary making*. Tübingen: Max Niemeyer.

Howarth, P. A. (1998). Phraseology and second language proficiency. *Applied Linguistics, 19*(1), 24-44.

Hsu, J.-y. (2007). Lexical collocations and their relation to the online writing of Taiwanese college English majors and non-English majors. *Electronic Journal of Foreign Language Teaching, 4*(2), 192–209.

Huff, K., Powers, D. E., Kantor, R. N., Mollaun, P., Nissan, S., & Schedl, M. (2008). Prototyping a new test. In C. A. Chapelle, M. K. Enright & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 187-226). New York: Routledge.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hunston, S., & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Institute of International Education. (2012). Top 25 places of origin of international students, 2010/11-2011/12. Open Doors Report on International Educational Exchange. Retrieved from http://www.iie.org/opendoors

International TA Program. (2012). SPEAK/TEACH test validity. Retreived from http://www.grad-college.iastate.edu/speakteach/testing_information/validity.php

Jun, H., & Li, J. (2010). Factors in raters' perceptions of comprehensibility and accentedness. In J. M. Levis & K. LeVelle (Eds.), *Proceedings of the 1st Pronunciation in Second Language Learning and Teaching Conference* (pp. 53-66). Ames, IA: Iowa State University.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Westport, CT: Praeger.

Kane, M. T. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 39-64). Charlotte, NC: Information Age.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kennedy, G. (2008). Phraseology and language pedagogy: Semantic preference associated with English verbs in the British National Corpus. In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 21-42).

Keshavarz, M. H., & Salimi, H. (2007). Collocational competence and cloze test performance: A study of Iranian EFL learners. *International Journal of Applied Linguistics, 17*(1), 81-92.

Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Kuiper, K. (1996). *Smooth talkers: The linguistic performance of auctioneers and sportscasters*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Laufer, B. (2011). The contribution of dictionary use to the production and retention of collocations in a second language. *International Journal of Lexicography, 24*(1), 29-49.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 61*(2), 647–672.

Leśniewska, J., & Witalisz, E. (2007). Cross-linguistic influence and acceptability judgments of L2 and L1 collocations: A study of advanced Polish learners of English. *EUROSLA Yearbook, 7*(1), 27-48.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levelt, W. J. M. (1999a). Models of word production. *Trends in Cognitive Sciences, 3*(6), 223-232.

Levelt, W. J. M. (1999b). Producing spoken language: A blueprint of the speaker. In C. M. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 83-122). Oxford: Oxford University Press.

Liao, Y., & Fukuya, Y. J. (2004). Avoidance of phrasal verbs: The case of Chinese learners of English. *Language Learning, 54*(2), 193-226.

Lin, P. M. S. (2010). The phonology of formulaic sequences: A review. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 174-193). London: Continuum.

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

Martyńska, M. (2004). Do English language learners know collocations? *Investigationes Linguisticae, XI*.

McGlone, M., Glucksberg, S., & Cacciari, C. (1994). Semantic productivity and idiom comprehension. *Discourse Processing, 17*(2), 167-190.

McNamara, T. (1996). *Measuring second language performance*. London: Addison Wesley Longman Limited.

Mei, J. J., (1999). *Dictionary of Modern Chinese Collocations*. Shanghai: Hanyu Dictionary Press.

MeritTrac. (2015). SpeechTRAC. Retrieved from http://www.merittrac.com/innovation-technology/products/speechtrac/

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012-1027.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Michou, A. & Seretan, V. (2009). A tool for multi-word expression extraction in Modern Greek using syntactic parsing. In *Proceedings of the Demonstrations Session at EACL 2009* (pp. 45-48), Athens, Greece. Retrieved from http://www.latl.unige.ch/personal/vseretan/publ/Michou_Seretan_2009.pdf

Millar, N. (2011). The processing of malformed formulaic language. *Applied Linguistics, 32*(2), 129-148.

Molavi, A., Koosha, M., & Hosseini, H. (2014). A comparative corpus-based analysis of lexical collocations used in EFL textbooks. *Latin American Journal of Content and Language Integrated Learning, 7*(1), 66-81.

Moon, R. (2008). Sinclair, phraseology, and lexicography. *International Journal of Lexicography, 21*(3), 243-254.

Moder, C. L., & Halleck, G. B. (2009). Planes, politics and oral proficiency: Testing international air traffic controllers. *Australian Review of Applied Linguistics, 32*(3), 25.21-25.16.

Moder, C. L., & Halleck, G. B. (2012). Designing language tests for specific social uses. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 137-149). London: Routledge.

Myers, J. L., & Chang, S.-F. (2009). A multiple-strategy-based approach to word and collocation acquisition. *International Review of Applied Linguistics in Language Teaching, 47*, 179-207.

Namei, S. (2004). Bilingual lexical development: A Persian–Swedish word association study. *International Journal of Applied Linguistics, 14*(3), 363-388.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics, 24*(2), 223-242.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.

Nesselhauf, N. (2009). Co-selection phenomena across New Englishes: Parallels (and differences) to foreign learner varieties. *English World-Wide, 30*(1), 1-26.

Nizonkiza, D. (2011). The relationship between lexical competence, collocational competence, and second language proficiency. *English Text Construction, 4*(1), 113–146.

Palmer, H. E. (1933). *Second interim report on English collocations*. Tokyo: Kaitakusha.

Pap, A. (1953). Reduction-sentences and open concepts. *Methodos, 5*, 3-30.

Paradis, M. (1981). Neurolinguistic organization of a bilingual's two languages. *The LACUS forum, 7*, 486-494.

Paradis, M. (2004). *A Neurolinguistic Theory of Bilingualism*. Amsterdam: John Benjamins.

Partington, A. (1998). *Patterns and meanings*. Amsterdam: John Benjamins.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191-226). London: Longman.

Piao, S. S., Rayson, P., Archer, D., & McEnery, T. (2005). Comparing and combining a semantic tagger and a statistical tool for MWE extraction. *Computer Speech & Language, 19*(4), 378-397.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences, 11*(3), 105-110.

Powell, B. (2012). *ETS reports the largest number of Chinese TOEFL® test takers in history.* Princeton, NJ: ETS.

Rashid, R. (2012, November). *Speech recognition breakthrough for the spoken, translated word*. [Video file]. Retrieved from https://www.youtube.com/watch?v=Nu-nlQqFCKg

Read, J., & Nation, P. (2004). Measurement of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing, and use* (pp. 23-35). Amsterdam: John Benjamins.

Renouf, A., & Banerjee, J. (2007). Lexical repulsion between sense-related pairs. *International Journal of Corpus Linguistics, 12*(3), 415-443.

Renouf, A., & Banerjee, J. (2008). The phenomenon of lexical repulsion in text. *Lingvisticæ Investigationes, 31*(2), 213-225.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*(1), 27-57.

Rossiter, M. J., Derwing, T. M., Manimtim, L. G., & Thomson, R. I. (2010). Oral fluency: The neglected component in the communicative language classroom. *Canadian Modern Language Review, 66*(4), 583-606.

Russell, B. (1915). *Our knowledge of the external world as a field for scientific method in philosophy*. Chicago: Open Court.

Schmidgall, J. E. (2013). *Modeling speaker proficiency, comprehensibility, and perceived competence in a language use domain* (Unpublished doctoral dissertation). University of California, Los Angeles.

Schmitt, N. (1998). Measuring collocational knowledge: Key issues and an experimental assessment procedure. *I.T.L. Review of Applied Linguistics, 119-120*, 27-47.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. New York, NY: Palgrave Macmillan.

Schmitt, N., & Carter, R. (2004). Formulaic sequences in action: An introduction. In N. Schmitt (Ed.), *Formulaic sequences* (pp. 1-22). Amsterdam: John Benjamins.

Schmitt, N., & Underwood, G. (2004). Exploring the processing of formulaic sequences through a self-paced reading task. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing, and use* (pp. 173-189). Amsterdam: John Benjamins.

Seretan, V. & Wehrli, E. (2006). Accurate collocation extraction using a multilingual parser. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (pp. 953–960), Sydney, Australia.

Shih, R. H.-H. (2000). *Collocation deficiency in a learner corpus of English: From an overuse perspective.* Paper presented at the the 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong.

Shohamy, E. G. (1984). Does the testing method make a difference? The case of reading. *Language Testing, 1*(2), 147-170.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J., & Moon, R. (1989). Forward. In J. Sinclair, P. Hanks, R. Moon, S. Bullon, R. Krishnamurthy, E. Pollard, D. Yuill, T. Lane, S. Smith, J. Brown, A. Capel, L. Heaslip & D. Williamson (Eds.), *Collins cobuild dictionary of phrasal verbs* (pp. IV-VI). Glasgow: Williams Collins Sons.

Siyanova, A., & Schmitt, N. (2007). Native and nonnative use of multi-word vs. one-word verbs. *International Review of Applied Linguistics in Language Teaching, 45*(2), 119-139.

Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *The Canadian Modern Language Review, 64*(3), 429-458.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510-532.

Skripnikova, I. (2012). The main difficulties when studying Russian verbs of motion in a figurative meaning. *Open Journal of Modern Linguistics*, *2*(4), 147-150.

Snow, C. E. (2010). Academic language and the challenge of reading for learning about science. *Science*, *328*(5977), 450-452.

Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics, 8*(2), 209-243.

Sugiura, M. (2002). Collocational knowledge of L2 learners of English: A case study of Japanese learners. In T. Saito, J. Nakamura & S. Yamazaki (Eds.), *English Corpus Linguistics in Japan* (pp. 303-323). Amsterdam: Rodopi.

Sung, J. (2003). *English lexical collocations and their relation to spoken fluency of adult non-native speakers* (Unpublished doctoral dissertation). Indiana University of Pennsylvania, Indiana, Pennsylvania.

Swain, M., Huang, L.-S., Barkaoui, K., Brooks, L., & Lapkin, S. (2009). *The speaking section of theTOEFL iBT™ (SSTiBT): Test-takers' reported strategic behaviors (ETS Research Rep. No. RR-09-30).* Princeton, NJ: Educational Testing Service.

Tabachnick, B. G., & Fidell, L. S. (2005). *Using multivariate statistics* (5th ed.). Boston: Pearson.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 560-619). Washington, D.C.: American Council in Education.

Tomarken, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, *99*(1), 90-99.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics, 17*(1), 84-119.

Tyler, A. (1992). Discourse structure and the perception of incoherence in international teaching assistants' spoken discourse. *TESOL Quarterly*, *26*(4), 713-729.

Underwood, G., Schmitt, N., & Galpin, A. (2004). The eyes have it: An eye-movement study into the processing of formulaic sequences. In N. Schmitt (Ed.), *Formulaic Sequences: Acquisition, processing and use* (pp. 155-172). Amsterdam: John Benjamins.

Van Lancker, D., & Canter, G. J. (1981). Idiomatic versus literal interpretations of ditropically ambiguous sentences. *Journal of speech, language, and hearing research, 24*(1), 64-69.

Van Lancker, D., Canter, G. J., & Terbeek, D. (1981). Disambiguation of ditropic sentences: Acoustic and phonetic cues. *Journal of Speech, Language, and Hearing Research, 24*(3), 330-335.

Voss, E. (2012). *A validity argument for score meaning of a computer-based ESL academic collocational ability test based on a corpus-driven approach to test design (*Unpublished doctoral dissertation). Iowa State University, Ames, IA.

Wang, L., Eignor, D., & Enright, M. K. (2008). A final analysis. In C. A. Chapelle, M. K. Enright & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 259-318). New York: Routledge.

Wang, X., Evanini, K., & Zechner, K. (2013). Coherence modeling for the automated assessment of spontaneous spoken responses. *Proceedings of NAACL-HLT 2013*, 814-819.

Wang, Z., Zechner, K., & Sun, Y. (2015, April). *Monitoring the performance of human and automated scores for spoken responses*. Paper presented at the 2015 American Educational Research Association (AERA) Annual Meeting, Chicago, IL.

Webb, S., & Kagimoto, E. (2011). Learning collocations: Do the number of collocates, position of the node word, and synonymy affect learning? *Applied Linguistics, 32*(3), 259–276.

Wehrli, E., Seretan, V., Nerima, L., & Russo, L. (2009). Collocations in a rule-based MT system: A case study evaluation of their translation adequacy. In *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation* (pp. 128–135), Barcelona, Spain. Retrieved from http://www.latl.unige.ch/personal/vseretan/publ/EAMT2009_EW_VS_LN_LR.pdf.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing, 27*(3), 335-353.

Whitehead, A. N., & Russell, B. (1910). *Principia mathematica* (Vol. I). Cambridge: Cambridge University Press.

Whitsitt, S. (2005). A critique of the concept of semantic prosody. *International Journal of Corpus Linguistics, 10*(3), 283-305.

Williamson, D. M., Bennett, R. E., Lazer, S., Bernstein, J., Foltz, P. W., Landauer, T. K., . . . Sweeney, K. (2010). *Automated scoring for the assessment of Common Core Standards.* Retrieved from https://www.ets.org/s/commonassessments/pdf/AutomatedScoringAssessCommonCoreStandards.pdf

Williamson, D. M., Bejar, I. I., & Mislevy, R. J. (2006). Automated scoring of complex tasks in computer-based testing: An introduction. In D. M. Williamson, I. I. Bejar & R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 1-14). Mahwah, NJ: Lawrence Erlbaum Associates.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2-13.

Wolf, M. K., Everson, P., Lopez, A., Hauck, M., Pooler, E., & Wang, J. (2014). Building a framework for a next-generation English language proficiency assessment system. *ETS Research Report Series, 2014*(2), 1-48.

Wolter, B. (2001). Comparing the L1 and L2 mental lexicon. *Studies in Second Language Acquisition, 23*, 41-69.

Wolter, B. (2006). Lexical network structures and L2 vocabulary acquisition: The role of L1 lexical/conceptual knowledge. *Applied Linguistics, 27*(4), 741-747.

Wolter, B., & Gyllstad, H. (2011). Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Applied Linguistics, 32*(4), 430-449.

Wolter, B., & Gyllstad, H. (2013). Frequency of input and L2 collocational processing: A comparison of congruent and incongruent collocations. *Studies in Second Language Acquisition, 35*(3), 451-482.

Wood, D. (2006). Uses and functions of formulaic sequences in second language speech: An exploration of the foundations of fluency. *Canadian Modern Language Review, 63*(1), 13-33.

Wood, D. (2007). Mandarin Chinese speakers in a study abroad context: Does acquisition of formulaic sequences facilitate fluent speech in English? *The East Asian Learner, 3*(2), 43-62.

Wood, D. (2009). Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: A case study. *Canadian Journal of Applied Linguistics, 12*(1), 39-57.

Wood, D. (2010). *Formulaic language and second language speech fluency: Background, evidence and classroom applications*. London: Continuum.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A., & Fitzpatrick, T. (2010). Pushing learners to the extreme: the artificial use of prefabricated material in conversation. *Innovation in Language Learning and Teaching, 4*(1), 37-52.

Wr ́obel, A. (2011). Formulaicity vs. fluency and accuracy in using English as a foreign language. In M. Pawlak, E. Waniek-Klimczak & J. Majer (Eds.), *Speaking and instructed foreign language acquisition* (pp. 55-65). Tonawanda, NY: Multilingual Matters.

Xi, X. (2008). What and how much evidence do we need? Critical considerations in validating an automated scoring system. In C. Chapelle, Y. R. Chung & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 102-114). Ames, IA: Iowa State University.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing, 27*(3), 291-300.

Xi, X. (2012). Validity in the automated scoring of performance tests. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 438-451). London: Routledge.

Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). *Automated scoring of spontaneous speech using SpeechRaterSM v1.0* (ETS research report No. RR-08-62). Princeton, NJ: Educational Testing Service.

Xi, X., Higgins, D., Zechner, K., & Williamson, D. (2012). A comparison of two scoring methods for an automated speech scoring system. *Language Testing, 29*(3), 371-394.

Xu, J. (2010). Using multimedia vocabulary annotations in L2 reading and listening activities. *CALICO Journal*, *27*(2), 311-327.

Xu, J. (forthcoming). Decipher the validity jargon for language testers. *Language Testing*.

Xu, J., & Xi, X. (2010). *Comparing human and machine judgments of collocations and relating them to speaking proficiency*. Paper presented at the 32nd Language Testing Research Colloquium, University of Cambridge, UK.

Yasuda, S. (2010). Learning phrasal verbs through conceptual metaphors: A case of Japanese EFL learners. *TESOL Quarterly*, *44*(2), 250-273.

Yoon, S.-Y., & Bhat, S. (2012). Assessment of ESL learners' syntactic competence based on similarity measures. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 600-608.

Yu, D., & Deng, L. (2014). *Automatic speech recognition: A deep learning approach*. New York: Springer.

Zareva, A. (2007). Structure of the second language mental lexicon: how does it compare to native speakers' lexical organization? *Second Language Research, 23*(3), 123-153.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication, 51*, 883-895.

Zhang, X. (1993). *English collocations and their effect on the writing of native and non-native college freshmen* (Unpublished doctoral dissertation). Indiana University of Pennsylvania, Indiana, Pennsylvania.

Zughoul, M. R., & Abdul-Fattah, H. (2003). Translational collocational strategies of Arab learners of English: A study in lexical semantics. *Babel, 49*(1), 59-81.

APPENDIX A: INSTITUTIONAL REVIEW BOARD APPROVAL

IOWA STATE UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Institutional Review Board
Office for Responsible Resear
Vice President for Research
1138 Pearson Hall
Ames, Iowa 50011-2207
515 294-4566
FAX 515 294-4267

Date: 11/5/2010

To: Jing Xu
1137 Pearson Hall

CC: Dr. Carol A Chapelle
339 Ross Hall

From: Office for Responsible Research

Title: Collocations in Spontaneous Non-Native English Speech

IRB Num: 10-461

Submission Type: New **Exemption Date:** 11/4/2010

The project referenced above has undergone review by the Institutional Review Board (IRB) and has been declared exempt from the requirements of the human subject protections regulations as described in 45 CFR 46.101(b). The IRB determination of exemption means that:

- **You do not need to submit an application for annual continuing review.**
- **You must carry out the research as proposed in the IRB application**, including obtaining and documenting informed consent if you have stated in your application that you will do so or if required by the IRB.
- **Any modification of this research should be submitted to the IRB on a Continuing Review and/or Modification form, prior to making <u>any</u> changes**, to determine if the project still meets the federal criteria for exemption. If it is determined that exemption is no longer warranted, then an IRB proposal will need to be submitted and approved before proceeding with data collection.

Please be sure to **use only the approved study materials** in your research, including the **recruitment materials and informed consent documents that have the IRB approval stamp**.

Please note that you must submit all research involving human participants for review by the IRB. **Only the IRB may make the determination of exemption**, even if you conduct a study in the future that is exactly like this study.